UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

DOTTORATO DI RICERCA IN

MODELLI E METODI PER L'ECONOMIA E IL MANAGEMENT
(ANALYTICS FOR ECONOMICS AND MANAGEMENT)

Settore Scientifico Disciplinare:
SECS-S/01 - Statistica
CICLO XXXVI

Statistical Methods for Reinforcement Learning Policy Comparison

Dottorando:
Dott. Romeo CASESA

Supervisor:
Prof. Maurizio CARPITA

Contents

List of Figures

List of Tables

UNIVERSITÀ DI BRESCIA

Abstract

Economia e Management
Analytics for Economics and Management

Doctor of Philosophy

Statistical Methods for Reinforcement Learning Policy Comparison

by Romeo CASESA

Inserito nel contesto dell'apprendimento per rinforzo (reinforcement learning - RL), questo lavoro analizza le implicazioni di decisioni statisticamente informate quando si confrontano diversi sistemi intelligenti, cioè diversi algoritmi o policy. A tal fine, vengono introdotti il concetto di statistical blocking, di intervalli di confidenza inferenziali (ICI) e il test di Skillings-Mack. Sebbene questi approcci non siano nuovi nella letteratura statistica, la loro applicazione al reinforcement learning risulta innovativa. L'uso del statistical blocking deriva dall'intuizione che la classificazione delle policy avviene spesso sulla base di una moltitudine di compiti o task: questa fonte aggiuntiva di variabilità può essere eliminata, risultando in test più potenti. Ciò viene dimostrato tramite la procedura di test Skillings-Mack applicata prima a un set di dati campione della letteratura sul reinforcement learning e confrontata poi, attraverso dati sintetici, con altri metodi correntemente utilizzati in letteratura per il confronto delle policy. I risultati ottenuti mostrano come il test di Skillings-Mack risulti migliore rispetto ai metodi attualmente utilizzati in letteratura, fornendo risultati statisticamente significativi per campioni di dimensioni inferiori; inoltre, la procedura supera gli altri metodi quando vengono confrontati task le cui medie risultino differenti. Proponiamo inoltre l'uso di intervalli di confidenza inferenziali nel campo della RL. Questi intervalli di confidenza sono calcolati sulla base di un fattore di correzione che ridimensiona gli intervalli di confidenza per permettere la sovrapposizione solo quando la differenza tra due policy non è statisticamente significativa. È introdotto un nuovo fattore di conversione per gli ICI che risulta non singolare anche quando i due intervalli di confidenza sono molto vicini.

UNIVERSITÀ DI BRESCIA

Abstract

Economia e Management
Analytics for Economics and Management

Doctor of Philosophy

Statistical Methods for Reinforcement Learning Policy Comparison

by Romeo CASESA

This work departs from the Reinforcement Learning (RL) setting and analyzes the implications of statistically informed decisions when comparing different intelligent systems, i.e. different algorithms or policies. To this aim, we introduce the concepts of statistical blocking, inferential confidence intervals (ICIs) and the Skillings-Mack test. Although these approaches are not new within the statistical literature, their applications to reinforcement learning is innovative.

The use of statistical blocking stems from the intuition that policy classification is often performed based on a multitude of tasks: this added source of variability can be removed with statistical blocking, leading to more powerful tests. This is shown with the Skillings-Mack test procedure which is applied on a sample dataset from the reinforcement learning literature and compared, through synthetic data, against other state of the art policy comparison methods.
Our results show that the Skillings-Mack test performs better than currently available state of the art methods providing statistically significant results for lower sample sizes; in addition, the procedure outperforms other methods when tasks with different mean scores are compared. We further propose the use of inferential confidence intervals within the field of RL. These confidence intervals are calculated based on a correction factor which scales the confidence intervals to allow overlap only when two policies are not statistically significantly different; in this way we allow an inference "by-eye" approach. A novel correction factor for ICIs is introduces which is well behaved even when the two confidence intervals are very close.

Acknowledgements

This work is part of an Industrial PhD Program with Apogeo Space s.r.l.
I want to thank the company for giving me the trust and freedom in pursuing the research direction I have chosen.

This work could not have been possible without the support of my friends and family who stood with me through these years; from my wife, Lorena, who spent with me most of the time dedicated to writing this thesis and has been my ever-present companion for long nights of work; from my old family who pushed me through this effort; from my old and new friends, those I met in Brescia and those I know since school; from my colleagues at Apogeo and at the University with whom we shared stories, successes and nights out.

Finally, I am extremely grateful to my supervisor Maurizio who tirelessly beared with me for all of this time.

Introduction

When approaching for the first time a new task, we, as humans, take a series of actions or decisions based on what we think leads to the best final outcome. Often this process is iterative, as we naturally make mistakes and learn from them. This concept is the central idea behind Reinforcement Learning (RL) which, according to Sutton and Barto [63], can be defined as follows:

> Reinforcement Learning is learning what to do - how to map situations to actions - so as to maximise a numerical reward signal.

The quote naturally leads to the idea of RL being close to the Data Science and Machine Learning research areas. However, this is not the only field associated with reinforcement learning: as the example at the start of this chapter highlights, RL has profound implications within the realm of psychology. This idea, which was originally present in the RL literature, introduces a philosophical and ethical perspective which enriches the implications of the RL research.
Although the setting of this work is within the RL literature, our focus is on statistical methods which can be applied in general and may benefit other areas of research too, such as Machine Learning.

This work starts with chapter 1 introducing the key concepts from the field of RL and performing a bibliometric analysis in order to uncover literature trends, major sources and authors.
In chapter 2, the implications of statistically informed decisions are analyzed within the setting of RL policy comparison, that is within the framework of RL algorithms. This leads to the innovations brought by this thesis: we criticise the common approach to RL task aggregation and propose two methods which help streamline the comparison between policies, which, to the author's knowledge, have never been used in RL.

1. The Skillings-Mack non parametric test procedure for the difference between two policies;

2. The inferential confidence intervals (ICIs) for an inferential procedure based on confidence intervals overlap.

These novel approaches allow to properly tackle the multiplicity of different tasks and to perform inference "by eye" through confidence intervals.

The use of the Skillings-Mack test further drives the introduction of statistical blocking within the field; this approach is proposed in order to remove variability resulting from the presence of multiple tasks (i.e. blocks). Indeed

the comparison between policies is often not straightforward because of the complex data distribution and the multiplicity of tasks against which each policy is tested. This concept and the Skillings-Mack test are discussed in chapter 2 where this procedure is applied to an available dataset from literature and show better performances than available state of the art methods.
This result is further analyzed through the use of simulated data in chapter 3 where multiple tasks and distributions are tested with increasing complexity to show performance of the Skillings-Mack test and other state of the art methods. Notably, the Skillings-Mack test performs better than other state of the art comparison methods especially under the critical scenario of different tasks having distributions with different means. The Skillings-Mack approach should therefore be preferred when a diverse set of benchmark tasks is being tested.

Chapter 2 also introduces inferential confidence intervals within the field of RL and demonstrate their use on a two-policy comparison. Our implementation of inferential CIs uses the Skillings-Mack test but the procedure holds in general for any two sample test; the calculated correction factor allows to scale descriptive CIs so to allow overlap only when the underlying test's null hypothesis is not rejected.
We further introduce a novel modification for the correction factor of ICIs: the modified ICIs share the same properties of the original implementation but apply a correction which scales with the difference between the means and is therefore well behaved even when the two means to be compared are very close.

For both approaches a custom developed Python code is available in appendix B.

Chapter 1

Reinforcement Learning Review

1.1 Introduction to RL

This work focuses on the modern RL; although psychology and animal trial-and-error learning research can be considered as a first step towards RL [63], we do not discuss these in this work. Modern RL theory is rooted in three intertwined research areas, namely optimal control, dynamic programming and statistical learning. Historically these research areas have been rather separate although the mathematical framework used is similar. Starting from the late '80s with the work of Watkins and Werbos [73, 74] the different research have merged and the modern framework of RL started to emerge.
It was until the late 2000's that RL suffered from limited applicability due to the difficulty to expand to high-dimensional scenarios and the need to handcraft effective state representations [63]. The emergence of Deep Statistical Learning addressed these challenges and allowed famous success stories such as RL algorithms playing the games of Chess, Go [58] or Starcraft [71, 70] at Grandmaster level, correctly simulating the folding of proteins [56] and controlling plasma reaction [37].

The term "reinforcement learning" has been used in literature to indicate multiple concepts. Although its initial use is linked to the field of animal psychology, in recent years the term refers mainly to two concepts: 1) the set of problems faced by an agent interacting by trial-and-error with a dynamic environment and 2) a set of methodologies which have been developed to solve these problems [38].
To outline the RL problem, the concept of agent and environment are key. The environment is the setting within which interaction occurs; environments are dynamic, they change in time, react to our actions and provide feedback; environments can be real or virtual and even simulated. The agent represents the other side of the system; the agent acts within the environment following certain 'rules'; it receives feedback from the environment and decides how to act. The interaction between agent and environment is the key of RL: through interaction loop and the corresponding feedback, knowledge is provided to the agent.

This concept is represented in figure 1 which shows the agent and the environment and their interactions. At each time t the agent and the environment state are represented by a variable s: this state can be either completely

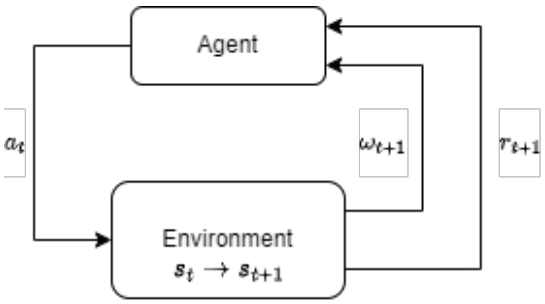


FIGURE 1: Agent and Environment interaction in Reinforcement Learning, adapted from [27]

known to the agent or partially known through an observation ω which is the information available to the agent. The agent performs an action a_t and receives a feedback (i.e. reward) r_{+1}. The interaction between agent and environment generates a sequence or trajectory as follows:

$$\omega_0(s_0), a_0, \omega_1(s_1), r_1, a_1 \ldots \tag{1.1}$$

By convention we represent together the action at time t and the following state s_{t+1} with its observation ω_{t+1} and reward r_{t+1}. Throughout this interaction the agent learns by trial and error the optimal behaviour which maximises reward [38]. Mathematically, this defines a discrete time stochastic process where an agent interacts with its environment and receives feedback from it [1]. Without interaction between environment and agent, we do not have a properly defined RL problem and instead fall in the category of "multi armed bandit" [63].

The learning process is thus centered around the reward, which encodes good and bad behaviour or outcome; with respect to supervised learning, where the agent (algorithm) is provided with the "right" answer (i.e. the dataset), in RL the aim is to learn a policy that acts differently from what is observed within the training dataset: in other words, the agent is allowed to find the best strategy [42].

1.1.1 RL Definitions

This section defines the RL terminology which will be used throughout this work. Given the close historical relation between the field of RL and Optimal and Dynamic control, it should come at no surprise that the two field share part of the terminology.

Environment The environment is the setting with which interaction occurs and which provides feedback to the agent in terms of state and reward. The environment can be real or simulated and may be stochastic. However the classical setting for RL requires the environment to be static or semi-static in

[1] The reader is referred to [63, 64] for a mathematical treatment of RL

the sense of not adapting its logic to the agent learning process. Examples of environments can be a cartpole, a videogame or a boardgame or even real-world scenarios such as a car driving on a road.

Agent The agent is the entity which actually interacts with the environment and takes decisions. Within the context of this work the scope of an agent will be limited to its behaviour and will ignore how the interaction with the environment is performed.

Reward This is the numerical quantity which is being maximized through the learning process and can be considered as the target or the return of the RL process. Contrarily to supervised learning, reward encodes a complex behaviour which is often difficult to separate in constituting elements; in facts, reward is often sparse and discrete. For example, playing a game may provide zero reward until the game is either lost (reward of -1) or won (reward of 1).

Policy A policy is the law by which the agent behaves and uniquely defines the agent behaviour under the same conditions. We will use this term interchangeably with the term *algorithm*.

Task A task is here defined as a specific RL environment. In general, it will be assumed that more than one task is performed by the same agent. Tasks can be part of a homogeneous setting: in this case, this group of tasks will be referred as a *benchmark*.

1.2 Bibliographic Analysis

The literature on RL is diverse and vast. To date, querying scientific databases such as SCOPUS or Clarivate Analytics' Web of Science returns more than 64000 scientific papers; a systematic review and literature analysis is therefore essential.
The scope of this section is to understand how the scientific literature has evolved through the years and to understand the major trends, contributors, journals and research groups [66, 23]. To this end, we follow an iterative approach: firstly within section 1.2.1, the rise of RL and its growth is highlighted; within section 1.2.3, a selection of articles is analysed; a selection of proceedings is analyzed in section 1.2.4; finally conclusions are drawn on the maturity of the field of study.

1.2.1 Overall Reinforcement Learning Literature

This section aims at demonstrating the growing interest in RL from the scientific community and providing some insight in the temporal evolution of the subject. This is possible by querying one of the well known scientific literature databases, namely Clarivate Analytics' Web of Science. We look for

a match on Title, abstract, author keywords, and "Keywords Plus[2]" of the following query:

$$reinforcement\ learning \tag{1.2}$$

The query results in roughly 64000 entries; this high number of results does not allow in-depth analysis on a personal computer and must be aggregated before use. Aggregated data is exported by Web of Science in table format and is at the base of the following discussion.
Based on this aggregated data, note the following:

- The annual scientific production is steadily increasing since early '90. This is shown in the top panel of figure 2 which summarizes the available annual scientific production. Notably, during the late '10s the trend is exponential[3].

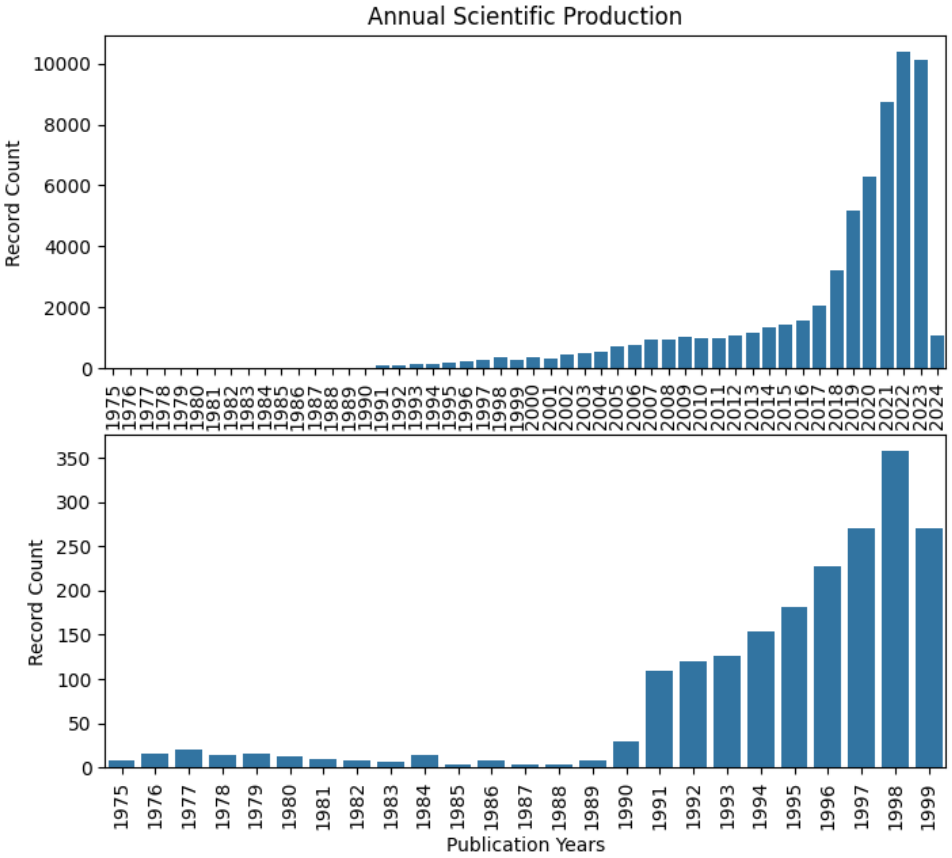
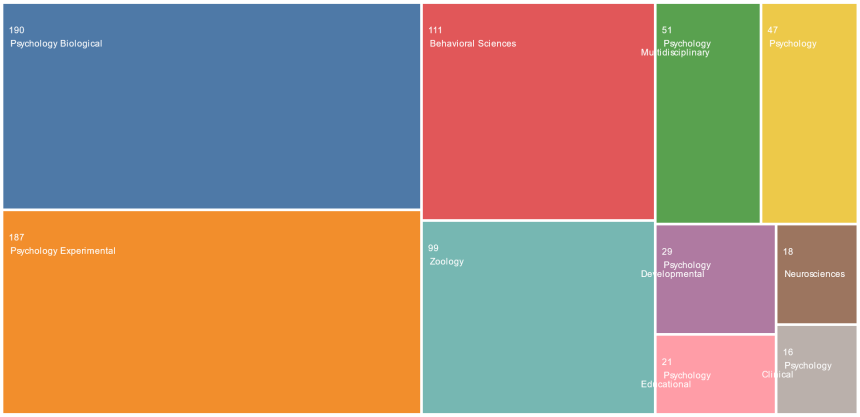


FIGURE 2: Annual Scientific production on RL

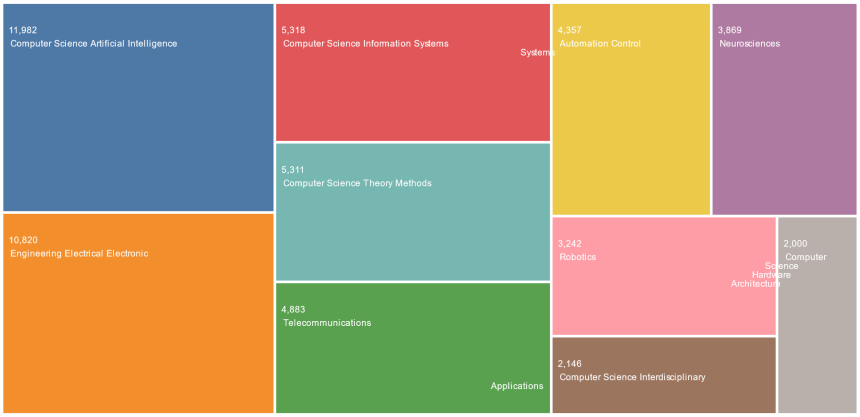
- The first entries are from 1975, but the scientific production is rather low and flat until the '90s with a yearly average article count of 13. This is shown in the lower panel of figure 2. Note the significant increase, starting from 1991.

[2]Keyword Plus are keywords automatically generated by a computer program, based on the keywords that appear on the document references and therefore not necessarily present in the document title [76]

[3]The year 2024 is, at the moment of writing, still ongoing and therefore shows a relatively low number of published documents.



(A) Before 1990



(B) After 2000

FIGURE 3: Major scientific production categories before 1990 and after 2000

- The above mentioned difference can be associated with a shift in the literature and the rise of RL as an independent field of study: whereas before the '90s the keywords "reinforcement learning" were associated mostly to the field of Biology and Psychology, after the '00s and until now, the same keywords are mostly associated with the field of Computer Science, Statistics and Artificial Intelligence. This is highlighted by the categorization provided by Web of Science's built-in analysis toolkit, shown in figure 3 where the top panel represents the major categories for the scientific production before 1990 and the bottom panel the major categories for the production after 2000.

1.2.2 Analysis of scientific production

The previous subsection highlighted the rise of RL and the growing interest of the scientific community; the analysis was however limited by the broad query and the large amount of resulting documents; in addition, the analysis did not capture correctly the topic of this work which focuses on statistical

methods for RL policy comparison. This subsection focuses on a more specific query on which to perform a bibliometric analysis.

To perform an in depth review of the scientific production, we make use of the R package *bibliometrix*, an open-source utility aimed at performing bibliometric analysis and science mapping [7]. With respect to section 1.2.1, the query was refined to make analysis possible on a personal computer.
The data was gathered from Clarivate Analytics' Web of Science: the use of this database allows a compromise between literature coverage and ease of analyse. The database's coverage has been shown to be lower than currently available alternatives such as Google Scholar [47] but it provides better selectivity [72] and offers tools for data extraction and analysis which are essential for the coming work.

The database query used is the following:

$$\begin{aligned} & AB = reinforcement\ learning \\ \textbf{and } & (AB = statistical \textbf{ or } AB = statistic \textbf{ or } AB = p - value) \\ \textbf{and } & WC = ComputerScience, ArtificialIntelligence \end{aligned} \tag{1.3}$$

Note the use of different categories within the query; for keywords *reinforcement learning*, *statistical*, *statistic* and *p-value* the category "AB" is specified which matches documents where the given keywords appears in the abstract; for keyword *Computer Science, Artificial Intelligence* the Web of Science's Category "WC" is specified which matches documents published in journals associated with this keyword (i.e. within this category).
Let us briefly comment on the individual keywords:

- *Reinforcement Learning*: the subject of this work; Up to this point the result would be the same as shown in the section 1.2.1. As discussed, this query results in more than 64000 results.
- *statistical* or *statistic* or *p-value*: restricts the query to results discussing statistical aspects of RL; in doing so, we are able to significantly reduce the number of entries from 64000 to nearly 1200 entries. The fact that less than 2% of papers uses this keyword in their abstract highlights a possible interesting research direction which is not broadly covered. Note that at least one of the three keywords must be present but not all of them need to be there for the filter to apply.
- *Computer Science, Artificial Intelligence*: removes contributions from research areas which are not directly involved with the generic RL setting. This is performed by making use of Web of Science's Subject Category; every entry in the database has at least one category (refer to [5] for the complete list) which is used to filter out papers proposing applications of the RL framework to specific field such as applied engineering and neuroscience; the inclusion of these specific application is beyond the scope of this work and is therefore filtered out.

In facts, if we were to leave these contribution within the dataset, the analysis of most cited articles and relevant journals, performed in later sections, would be polluted by the high number of citations and articles these applied communities generate, over-representing applied journals such as *IEEE Transactions on Vehicular Technology* or *IEEE Access*.

Notably, the Statistics and Probability category is not included as it accounts for only 20 results within this query. Considering the Computer Science, Artificial Intelligence category reduces the dataset to roughly 319 results.

- The query does not include wildcard characters as Web of Science automatically includes Lemmitization and Stemming to query results [6]. Lemmitization reduces inflected forms of a word to their lexical root, whereas stemming removes suffixes such as -ing and -es; we therefore expect the keyword *statistics* to be included when querying for keyword *statistic*.

- We did not include keyword "test" as this is a very generic word used in many areas and therefore leads to an excessively wide query. In facts, when adding this keyword to our query we obtain over 2000 results, which is nearly an order of magnitude higher than query 1.3.

The query 1.3 resulted in a total of 319 documents spanning the time frame 1992-2024. The time-span is consistent with the rise of modern RL as a discipline. In a similar fashion to what discussed in section 1.2.1, the annual scientific production on this reduced dataset is shown in figure 4, proving a positive annual growth rate of roughly 7%. The trend is comparable with what shown in figure 2 with a strong increase in scientific production from 2019 onwards.

The contributions are distributed over 212 different sources (Journals, Books etc..) divided into 188 proceedings (including pre-print articles) and 131 articles (including book chapters and reviews). The analysis of both document types will be performed within section 1.2.3 and 1.2.4 for articles and proceedings respectively.

Dataset modification

Within the dataset downloaded by Web Of Science, some minor differences in namings have been found and corrected prior to analysis. These correction are aimed at avoiding pitfalls in bibliometric analysis which are due to the tool not recognizing mispelled, abbreviated or alternative names or duplication of affiliation. The following have been corrected:

- Naming homogenization of the journal *Advances in Neural Information Processing Systems* which could be found either as *Adv Neural Inform Pr* or abbreviated as *NeurIPS* or *NIPS*.

- Removed duplication of *University of California System*.

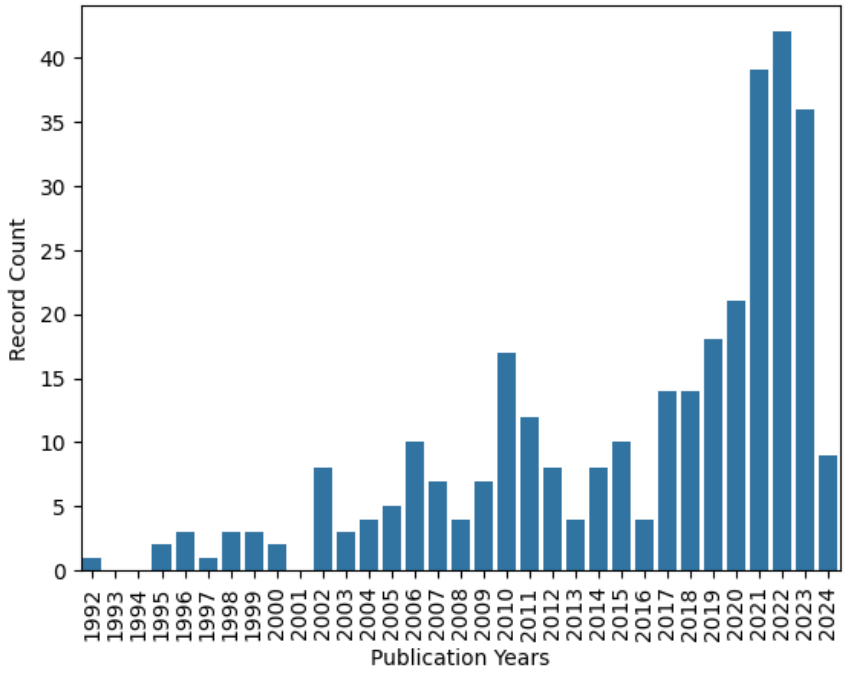


FIGURE 4: Annual scientific production on query 1.3

- Corrected reference to [Anonymous], 2010, Dynamic Programming
- Corrected reference to the online available pre-print of Sutton and Barto's *Reinforcement Learning: An Introduction* which was showing as Anonymous.
- Removed information on conference year or volume in order to group all conferences together (e.g. the *international conference on machine learning vol 162*, became the *international conference on machine learning*). This was applied to the international conference on machine learning and to the following conferences:
  - AAAI Conference On Artificial Intelligence
  - Advances In Neural Information Processing Systems
  - IEEE Conference On Intelligent Transportation Systems (ITSC)
  - IEEE International Conference On Development And Learning (ICDL)
  - IEEE International Joint Conference On Neural Networks (IJCNN)
  - European Conference On Artificial Intelligence (ECAI)
  - International Conference On Artificial Intelligence And Statistics (AISTATS)

1.2.3 Analysis of selected articles

This section focuses on a reduces dataset which includes only contributions from articles. The number of documents analyzed is reduced this way to 131 out of the 319 discussed in previous section.

Sources

The dataset includes 61 different sources. Because the dataset is limited to articles only, most of sources are journals. In order to appreciate the most relevant sources two metrics are analyzed:

- The most relevant source in terms of articles within the dataset; in other words, we count the number of articles published in each source within the dataset and list the top 10. The result is shown in table 1.1.

Sources	Articles
Journal Of Machine Learning Research	10
Expert Systems With Applications	7
Applied Intelligence	5
Applied Soft Computing	5
Engineering Applications Of Artificial Intelligence	5
Machine Learning	5
Neural Networks	5
Algorithms	5
IEEE Transactions on Neural Networks and Learning Systems	4
Knowledge-Based Systems	4

TABLE 1.1: Sources with more articles within the dataset

  Note the relatively small number of articles for each source. This is indicative of a rather sparse literature with no single major journal. A similar conclusion can also be drawn from the high number of sources in the dataset, nearly half the number of documents.
  The temporal dynamics of the 5 most relevant sources is shown in figure 5; notably we recognize the Journal *Machine Learning* as aggregating initial research in the field but being then surpassed by the *Journal of Machine Learning Research* which results the most relevant to date.

- The most relevant sources in terms of citations: in other words we count the number of times a source appears in a citation within the dataset. Table 1.2 shows the top 10 sources according this metric.
  As the number of references is generally large, the top journals get cited a lot more. Notably the first couple of entries in the list represent proceedings; this leads us to the suggestion that most of the research happens on proceeding papers. We further note the large difference in citations between the top two sources (i.e. *Arxiv* and *NeurIPS*) and all the other sources with the top two entries being cited nearly as much as the other 8 entries all together.
  The third entry is the *Journal of Machine Learning Research* which, aligns with the result from table 1.1.

Note the two metrics provided diverse results with some notable overlap; this can be explained by the fact most citations are proceedings but those

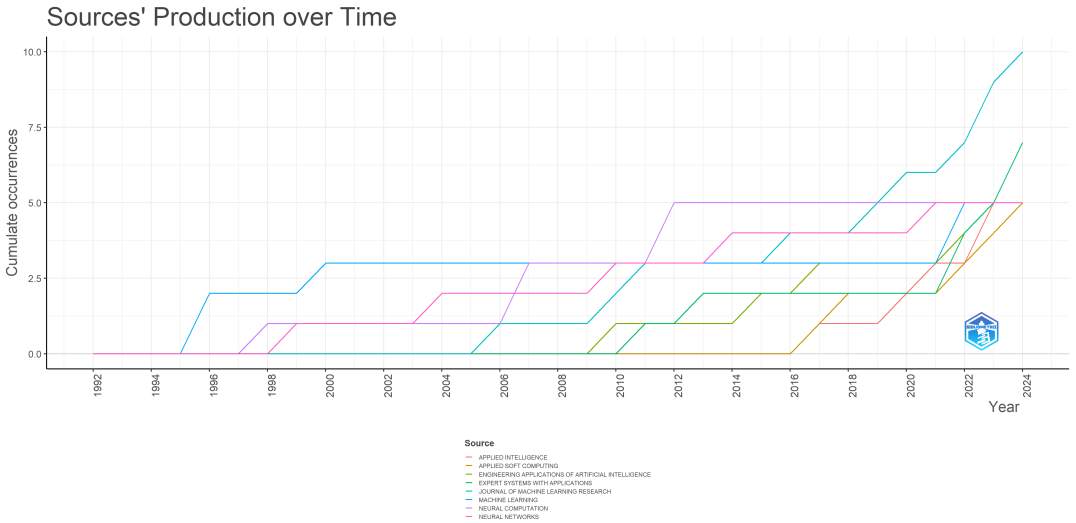


FIGURE 5: Source dynamics

Sources	Articles
Arxiv	261
Advances in Neural Information Processing Systems (NeurIPS)	258
Proceeding of Machine Learning Research	123
Journal of Machine Learning Research	113
Machine Learning	107
Nature	73
Expert Systems with Applications	64
Lecture Notes on Computer Science	56
Neural Computation	55
IEEE Transactions on Neural Networks and Learning Systems	51

TABLE 1.2: Sources being cited more often

articles have been filtered out in this section (refer to section 1.2.4 for an analysis on proceedings). The *Journal of Machine Learning Research* ranks high in both table 1.1 and 1.2. This journal can be considered a reference point for articles within the literature of interest.

Authors, Affiliations and Countries

The dataset includes 435 authors. In this subsection, the most relevant authors in the field are analyzed first; afterwards their geographic location is highlighted together with major affiliations.

Authors To understand which are the most relevant authors, different metrics are compared as follows:

- By number of published articles within the dataset. This provides an overview on most productive authors; clearly being more productive does not imply relevance per se. Resulting top 10 authors are listed in

table 1.3. We note that scholars ranked 2nd to 23rd all share the same number of published articles within the dataset; this made necessary a second level sorting based on the fractionalized number of articles [4].

Based on this list it is further possible to analyze the authors pro-

Authors	Articles	Articles Fractionalized
Wang L	3	0.78
Iwata K	2	1.33
Lemon O	2	1.25
Cho SY	2	0.83
Zheng L	2	0.83
Bhatnagar S	2	0.58
Mannor S	2	0.58
Prabuchandran KJ	2	0.58
Wang JJ	2	0.58
Young S	2	0.58

TABLE 1.3: Most productive authors within dataset

duction over time; this is shown in figure 6 which presents all authors with at least 2 published articles within the dataset. In the figure, the red lines indicate period of research activity, delimited by publications shown as dots; bigger dots indicate a higher number of published papers.

[4]Fractional authorship quantifies an individual author's contributions to a published set of papers, following the hypothesis of uniform contribution of all co-authors for each document [8]

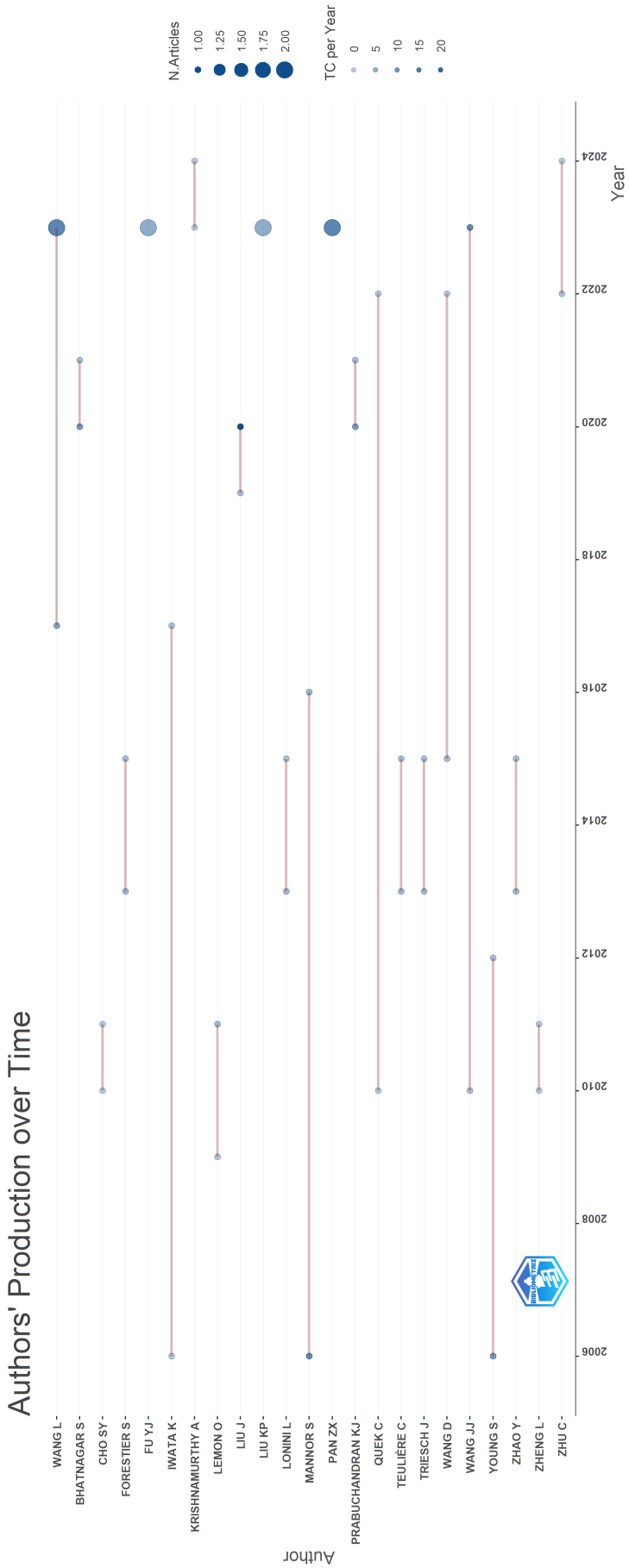


FIGURE 6: Author annual production over time for authors with at least 2 published articles within the dataset.

- By number of citations within the dataset (local citations). With the exception of *Young S*, all authors are only cited once or less within this dataset.

- By H-index; this index, also known as Hirsch index,is obtained by analysing the citations within the dataset and provides an objective function to score different scientists. Formally, we say an author has h-index H, if H of his/her N_p (i.e. the number of articles published) articles have received at least H citations each, and the rest $(N_p - H)$ articles have received no more than H citations [57]. The index tries to merge both impact and productivity into a single criterion. The resulting top 10 authors according this ranking are shown in table 1.4. Tied h-index scores are resolved by the total citation score.
  The result is similar to table 1.3.

Author	h index	g index	m index	TC	NP	PY start
Wang L	3	3	0.375	79	3	2017
Mannor S	2	2	0.105	264	2	2006
Young S	2	2	0.105	193	2	2006
Liu J	2	2	0.333	135	2	2019
Bhatnagar S	2	2	0.4	51	2	2020
Forestier S	2	2	0.167	51	2	2013
Lonini L	2	2	0.167	51	2	2013
Prabuchandran Kj	2	2	0.4	51	2	2020
Teulière C	2	2	0.167	51	2	2013
Triesch J	2	2	0.167	51	2	2013

TABLE 1.4: Authors with highest h-index within dataset. TC: total citations; NP: number publications; PY start: start of publication activity

The three metrics described are aligned in describing the sparsity of results: most authors have published only a couple of articles within this specific research. This conclusion can be drawn both by analyzing tables 1.3 and 1.4 and similarly from from figure 6 which highlights the sporadic nature of most contributions and their uniform distribution in time.
In addition no single author appears to stand-out in terms of relevance within the research with the three metrics providing different top authors; the only exception is author Wang Ling from Tsinghua University, which results first both following the total articles metrics and the h-index metric.

As suggested within section 1.2.2 this may indicate a field of research which is not mature yet where contributions are sparse and no single author stands out.

Geographic location We proceed by analyzing the geographic location in terms of affiliation and country.

By counting the occurrence of each affiliation within the dataset, it is possible to provide a metric for the most relevant affiliation. Note however this is biased towards the most productive authors or affiliations and does not take into consideration impact.
Table 1.5 shows the resulting top 10 affiliation based on this criterion. Notably the link between top authors and affiliation is rather weak. Following the h-index rank from table 1.4, we see that top author Wang Ling's affiliation (i.e. Tsinghua University) appears within the affiliation rank; this is not the case for other top authors from table 1.4 such as Iwata's affiliation (Hiroshima City University) or Lemon's affiliation (Heriot Watt University). Similarly according the ranking of table 1.3 scholars Mannon Shie's affiliation (Technion Israel Institute of Technology) and Young Steve's affiliation (University of Cambridge) do not appear within the rank of table 1.5 as well.

Affiliation	**Articles**
University Of California System	7
Nanyang Technological University	4
Tsinghua University	4
Agency For Science Technology And Research (A*STAR)	3
Harvard University	3
University Of Calgary	3
University Of London	3
Wuhan University	3
Zhengzhou University	3

TABLE 1.5: Top productive affiliations

Table 1.5 gives a preliminar view on which are the most active countries too; United Stated and Singapore appear to play a leading role.
This is highlighted by figure 7, which shows the most cited countries within the dataset. It should be noted that figure 7 represents absolute numbers and is therefore biased towards bigger countries and/or with a larger number of research institutions.

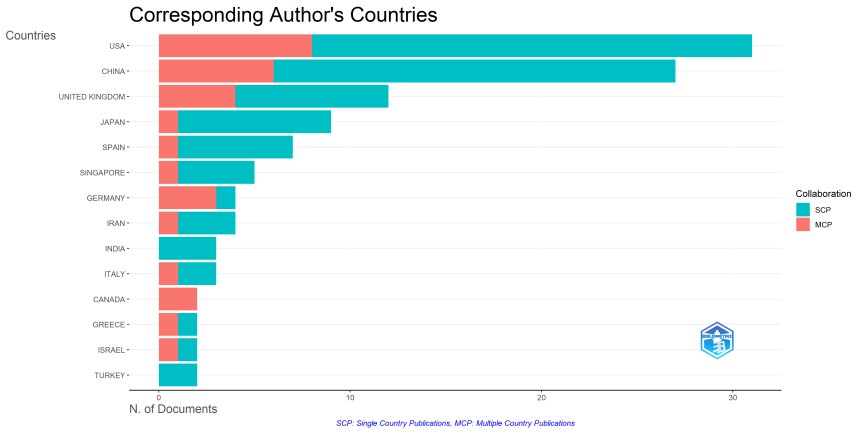
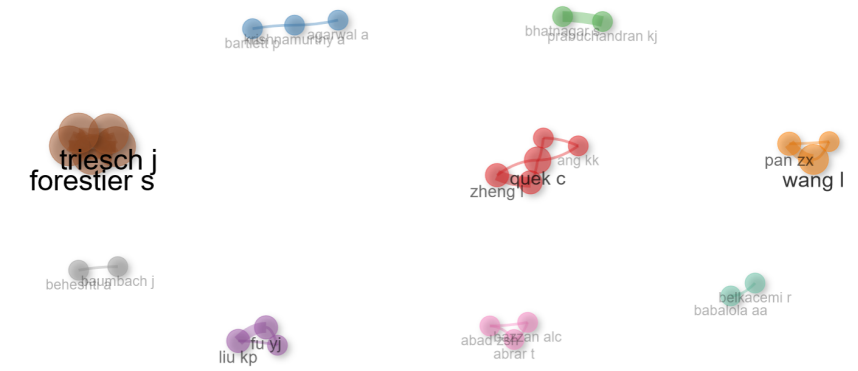
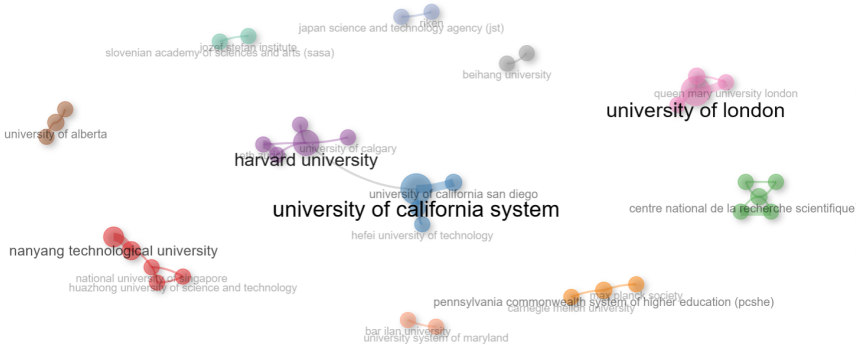


FIGURE 7: Most relevant countries

Research groups & collaborations In this paragraph the link between authors and universities is analysed in order to uncover the hidden groups of scholars, regular study groups and university clusters.



(A) Authors collaboration Network



(B) University collaboration network

FIGURE 8: Collaboration network highliting links between universities and authors working jointly

This information is visually contained in figure 8 which shows two collaboration networks, for authors in the top panel and for universities in the bottom one. The network highlights links between different nodes and clusters, grouped by color. The spacial layout follows the multidimensional scaling (MDS) plotting convention where distance is a measure for similarity of the nodes; in other words two nodes that are close one another are also more similar. Note that similar coloring between the top and bottom panels do not indicate a corresponding link but are casual.

From figure 8, wee note the presence of multiple small clusters mainly associated with single universities. With the exception of a link between Harvard University and the University of California System, no the different study group are separate. This may indicate the presence of silos within the community or a fragmentation of topics, meaning that through the query we grouped a multitude of fields of research which is not representative for a single field of research.

1.2.4 Analysis of selected proceedings

Whereas subsection 1.2.3 is focused on the analysis of scientific articles, this section concentrates on proceedings. Given the large number of proceedings within the query, the analysis of proceedings is performed separately from the research on articles which was the objective of the previous section. This division further allows to avoid polluting different sources and to highlight the contributions from major conferences.
The analysis is based on the dataset resulting from Web of Science through query 1.3 with the additional filter being applied to include only proceeding papers.

This results in 188 proceedings produced through the years 1995-2024. The trend is similar to what highlighted for articles production with a positive annual growth rate of roughly 9% until 2023[5]. Figure 9 details this trend on a year by year basis plotting both the proceeding and the article record count.
During the period 2017-2021 the scientific production increased exponentially peaking in 2021 and then dropping in the last two years. This result is aligned with what was shown in figure 4.

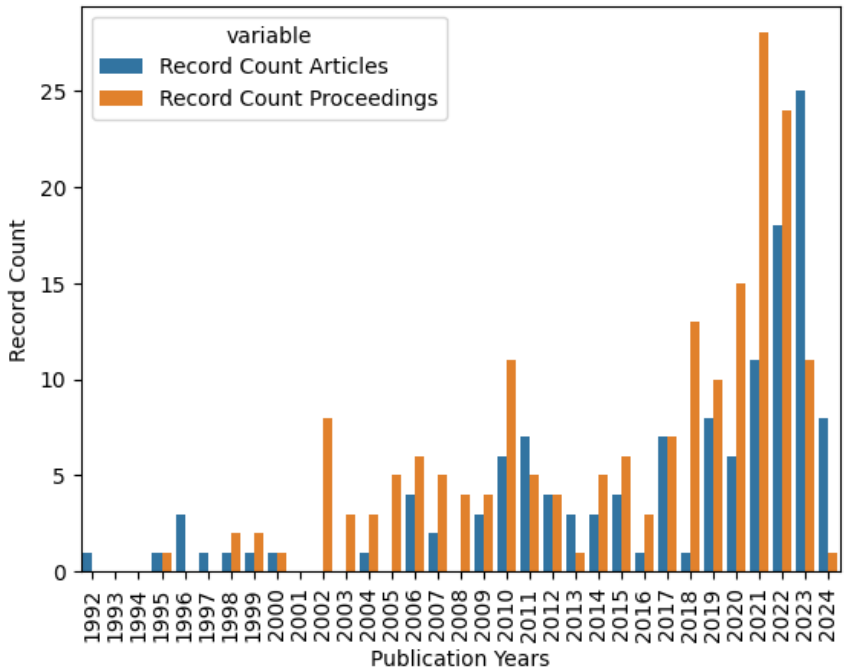


FIGURE 9: Annual proceeding and article production compared

Notably, the peak in proceedings production is, with respect to the article annual production, shifted a couple of years in advance; this can be expected

[5]Year 2024 has been removed from this calculation as it is not yet ended and the resulting number of documents is significantly lower than would be expected.

as many conferences (where proceedings are published) gather the scientific community and pave the way for future research in the field.

Sources and Authors

The analysis of proceeding highlights the direction the scientific community is moving and allows inferring major conferences; the latter is summarized in table 1.6 which shows the major sources of published proceedings. The table lists the number of times a source appears within the dataset, i.e. the amount of papers published in that source (within the dataset).

Sources	Articles
International Conference on Machine Learning	19
AAAI Conference on Artificial Intelligence	11
Advances In Neural Information Processing Systems	11
IEEE Conference on Intelligent Transportation Systems (ITSC)	6
IEEE International Conference on Development And Learning (ICDL)	4
IEEE International Joint Conference on Neural Networks (IJCNN)	4
48Th Annual Meeting of The Association For Computational Linguistics	3
European Conference on Artificial Intelligence (ECAI)	3
International Conference on Artificial Intelligence And Statistics (AISTATS)	3
Empirical Methods In Natural Language Generation	2

TABLE 1.6: Major conferences within the query

Notably, most of conferences focuses on the broader machine learning and artificial intelligence field; only one conference, namely AISTATS, actually specifically mentions statistics. Also there are three entries for IEEE conferences which, by its own definition, typically targets "engineering, computing, and technology information" [35].

According to the local citation criteria, the rank is shown in table 1.7; this result closely matches the one provided in table 1.2 and confirms the importance of Arxiv within the Artificial Intelligence and Machine Learning community [62].

When analyzing the author keywords from the query 1.3 we find that only 6 papers mention the term *statistical*, *statistic* or *p-value* within the article's author keywords: this list is provided in the table 1.8. We note the diversity of authors and sources similar to the results of previous sections.

The selection of papers targets the optimization of RL policies or its application to specific fields and is therefore considered not relevant to the scope of this work.

Sources	Articles
Arxiv	316
Advances In Neural Information Processing Systems	193
Proceeding of Machine Learning Research	178
AAAI Conference on Artificial Intelligence	72
Journal of Machine Learning Research	72
Machine Learning	71
Nature	63
Lecture Notes on Artificial Intelligence	47
Lecture Notes on Computer Science	46
Journal of Artificial Intelligence Research	41

TABLE 1.7: Most local cited sources of proceedings

1.2.5 Alternative database

The previous sections have been focused on query 1.3 at Web of Science database. Clearly, the results obtained are only as relevant and complete as the database from which we got the data from. We recognize this assumption to be critical as most of the cited references within the query refer to articles published on ArXiv (see table 1.2 and 1.7), which is not indexed by Clarivate Analytics' Web of Science [4].

Given this evidence, an attempt is made to compare top results articles with Google Scholar. The following query is performed:

$$reinforcement\ learning\ statistic* \tag{1.4}$$

Note the use of the wildcard character $*$ to indicate any word which starts with *statistic*. This query therefore includes both the term *statistical*, *statistic* and *statistics* used in previous sections. It is on the other hand not possible to limit search to specific categories as was done in section 1.2.2. Notably by running query 1.4, we are returned a total of 287000 results: this is vastly larger than what was available on Web of Science, even considering the generic query 1.2. However, Google Scholar does not support exporting of search results and meta-data analysis: therefore we focus on comparison of the top results from this query which are shown in table 1.9.

Interestingly, these results do present the keyword *statistic* or *statistics* within their title. We further note that two papers from author Colas C. are included within the table both published on ArXiv. In facts, nearly half of the articles within table 1.9 are published on ArXiv.

1.2.6 Conclusions

The analysis of bibliometric data demonstrated a field of study which is not mature yet with sparse contributions and the absence of dominating research. Different metrics do not agree on a rank of dominating authors: in fact most authors and affiliations only contributed with a couple of articles.

Authors	Title	Source
Pan, ZX et al.	A Learning-Based Multipopulation Evolutionary Optimization for Flexible Job Shop Scheduling Problem With Finite Transportation Resources	IEEE Transactions on Evolutionaty Computation
Ultes, S et al.	Domain-independent User Satisfaction Reward Estimation for Dialogue Policy Learning	Annual Conference of the International Speech Communication Association
Sarkar, S et al.	Reinforcement Learning for Pass Detection and Generation of Possession Statistics in Soccer	IEEE Transactions on Cognitive and Developmental Systems
Vergara, G et al.	Deep reinforcement learning applied to statistical arbitrage investment strategy on cryptomarket	Applied Soft Computing
Ghesu, FC et al.	Towards intelligent robust detection of anatomical structures in incomplete volumetric data	Medical Image Analysis

TABLE 1.8: Papers citing term *statistical*, *statistic* or *p-value* within their author keywords

The only notable exception is Colas C. which figured within Google Scholar's query as the only author using the keyword *statistic* within its articles and publishing more than one article. This insight provides evidence on the possible need to develop statistical research within the field of RL. However we shall note the limitations of the performed analysis: most of the cited papers are published on ArXiv and are not included within common databases such as Web of Science's. This may lead to the failure to recognize important contributions.
Finally we recognize two major sources for research within the field, namely *Advances in Neural Information Processing Systems* and the *Journal of Machine Learning Research*.

Authors	Title	Source
Rowland M. et al	Statistics and samples in distributional reinforcement learning	Machine Learning
Colas C. et al	A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms	ArXiv
Agarwal R. et al.	Deep reinforcement learning at the edge of the statistical precipice	Advances in neural information processing systems
Kane D. et al	Computational-statistical gap in reinforcement learning	Conference on Learning Theory. PMLR
Williams R.J. et al	Simple statistical gradient-following algorithms for connectionist reinforcement learning	Machine Learning
Colas C. et al	How many random seeds? statistical power analysis in deep reinforcement learning experiments	ArXiv
Chan S.C.Y et al	Measuring the reliability of reinforcement learning algorithms	ArXiv
Lahoudakis M.G. et al	Algorithm Selection using Reinforcement Learning.	ICML
Nguyen-Tang T. et al	Distributional reinforcement learning via moment matching	Proceedings of the AAAI Conference on Artificial Intelligence
Borkar V.	The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning	ArXiv

TABLE 1.9: Top papers by relevance from query 1.4

Chapter 2

Policy Comparison

The aim of this chapter is introducing the RL policy (or algorithm) comparison literature and proposing two novel approaches to policy comparison. The proposed methodologies are based on three key concepts:

- Statistical blocking: a methodology to harness data's structure to remove variability;
- Skillings-Mack's test procedure: a non-parametric hypothesis test for the difference of the means which makes use of statistical blocking;
- Modified Inferential Confidence Intervals: a novel computation for Inferential Confidence Intervals, which can be used to perform inference on the difference of the means.

Through this chapter we will formally define these concepts and perform a comparative analysis against results obtained in literature, specifically comparing with inference based on stratified bootstrap confidence intervals. Chapter 3 will extend these results to a controlled setting and demonstrate performances of the introduced methods against other state of the art methodologies such as the Welch, Yuen and t test.
Notably, the concepts of statistical blocking and the Skillings-Mack's test are not new within the statistical literature but, to the author's knowledge, constitute an innovation within the field of RL.

2.1 Policy Comparison in RL Literature

This section introduces the rationale which drives this work and presents the challenges faced.

We introduced in section 1.1 the concept of RL being both a set of problems and a methodology to solve these challenges. Not only are we concerned with improving our capability of solving complex tasks but also in developing solutions which generalize well to a multitude of tasks, that is in developing an *intelligent system*.

Through our search of an intelligent system we are naturally led to the definition of a set of tasks which we require our artificial intelligence to master; in other words, we strive to find methods which are capable of dealing

with different (and complex) tasks. Within the RL research community this translates into different agent's policies.

This definition of what is a - so to say - *good* artificial intelligence is, of course, limited. The issue of defining the boundary of intelligence only moved to the definition of a set of tasks which are believed should be accomplished by a *good* intelligence. This rather ethical question is left to the reader; in the following we will assume that a set of tasks, which represent a desirable benchmark for our intelligent system, is available.

We are finally led to the question of how do we compare two or more RL policies based on tasks which are dissimilar, may have different outputs and complexities; indeed section 2.1.2 will show that a plethora of different standard tasks is available. When few tasks are compared it is possible to display complete learning curves and this is done in some cases [49, 53]. The learning curve approach however becomes infeasible when multiple tasks are compared [44].

This question will be the focus of the coming discussion. In section 2.1.3 an attempt is made in summarizing methodologies used in literature for aggregate learning data and compare multiple policies in the framework of complex benchmark environments involving multiple tasks, such as the Atari 100k [39]. Section 2.2 and 2.3 describe the proposed approach which is then applied to the RL framework in section 2.5. The discussion will focus on offline policy evaluation [32] although a generalization to online learning is possible.

2.1.1 Research Question and Notation

Before moving further, there is the need to formalize the focus of the coming discussion.
Let n be the number of tasks available and let k be the number of different policies we want to compare. Given k policies $\pi_1, \dots, \pi_k$ the aim of this work is determining which performs better on a series of n tasks $\tau_1, \dots \tau_n$. When running policy π_1 on task τ_1 the end-score (i.e. the result of that run) can be considered a realization of the random variable X_{11}. In general for each i-task and j-policy combination we have a random variable X_{ij} which results in a sample of size c_{ij}, depending on the number of runs (i.e. simulation repetitions) performed. This is represented in table 2.1.

In general each random variable X_{ij} is not required to follow the same distribution; this is true both across tasks (same policy) and across policies (same task).

When performing a comparison between two policies, intuition guides us to the following null hypothesis:

$$H_0 : \ \pi_1 = \pi_2 \tag{2.1}$$

Tasks	Policies			
	Policy 1	Policy 2	...	Policy k
Task 1	X_{111} ... $X_{11c_{11}}$	X_{121} ... $X_{12c_{12}}$	...	X_{1k1} ... $X_{1kc_{1k}}$
Task 2	X_{211} ... $X_{21c_{21}}$	X_{221} ... $X_{22c_{22}}$	...	X_{2k1} ... $X_{2kc_{2k}}$
...	...	...	...	...
Task n	X_{n11} ... $X_{n1c_{n1}}$	X_{n22} ... $X_{n2c_{n2}}$	...	X_{nk1} ... $X_{nkc_{nk}}$

TABLE 2.1: Data notation

Which, in case of multiple policies, becomes:

$$H_0 : \pi_1 = \cdots = \pi_k \tag{2.2}$$

These hypotheses state that no difference exists between policies. Clearly, with respect to representing every policy-task comparison, the information within equations 2.1 and 2.2 is more synthetic but carries the complexities discussed at the beginning of this chapter.

The corresponding alternative hypotheses state that a statistically significant difference exist between the policies:

$$H_1 : \pi_1 \neq \pi_2 \tag{2.3}$$

Which, in case of multiple policies, becomes:

$$H_1 : \pi_1, \ldots, \pi_k \quad \text{not all equal} \tag{2.4}$$

Note that, whereas the result of rejecting hypothesis 2.1 is clear meaning that one of the two policies is better than the other, the result of rejecting the null hypothesis 2.2 is harder to interpret given that it is sufficient for one policy to be significantly better to reject.

2.1.2 RL benchmarks

Comparison of RL policies is a challenging tasks due to the stochastic nature [27] and the complexity [9] of the environment. In literature, different attempts have been made to standardize the comparison of RL policies. With respect to supervised learning, within RL the interaction with the environment does not allow the creation of static dataset; we therefore speak of benchmarks, i.e. environments with which the agent can interact.

Following early research influences from the dynamic programming and optimal control fields, early work on RL focused on classic control problems such as holding an inverted pendulum in equilibrium [10]; these problems lack however the complexity of real scenarios.
Board games were also extensively used; one of the major leaps forward in RL was due to Tesauro who proposed an algorithm which achieved master-level play at backgammon [65] and, more recently, the game of Go [58, 59].
Currently, the main benchmarks for RL are represented by videogames; these provide a complex scenario which is challenging for humans as well provided that often games have large observation/action spaces and generally require long planning horizons [27].
One of the widely used benchmarks environment is based on a set of 2600 Atari games [12] which provides a wide range of domain-independent tasks designed for RL and planning; this model was later adapted to reduce its computational requirements by Kaiser et al and is known as the Atari 100k [39]. The Atari 100k will be the baseline for the next sections.

2.1.3 State of the Art policy comparison

Through the current RL literature a multitude of different methods has been adopted to compare policies. Many researchers use statistical and common-sense techniques to assess policy performance differences with the majority of papers reporting only mean or median scores on a handful of samples [2]. Unless otherwise noted the evaluation is assumed to be performed on the simulation end-result (commonly referred as the *return* of a run), that is the score obtained by the specified policy on the specific task. This value encodes how well a policy performed on a certain task.

This section discusses a selection of methods used in literature and addresses for each of them some criticalities.

Point Estimates

The most widely adopted methodology for comparison of RL policies makes use of point estimates of policy performance; the comparison is based on a set of calculated metrics which, to date, is not standardized yet; the most common metrics used are [36]:

- Maximum return
- Standard deviation return
- Average return

other metrics have been proposed, including median scores and Inter-Quartile Mean (IQM) [2].
These metrics constitute point estimates of the corresponding population parameter and are commonly used to make inference about which policy performs better. As highlighted by [32], this leads to at least two sets of issues:

on one hand, some of these metrics are biased (e.g. maximum return and maximum average return [36]) which lead to unreliable comparisons; on the other hand, the uncertainty associated with the estimate is often ignored.
The latter point is key to the "reproducibility crisis" experienced in machine learning and RL and is linked to the high variability of end-results. This variability will be further discussed in section 2.1.4; however, the use of point estimates results critical when it leads to simple inferences based only on which policy has the highest metric, ignoring the data distribution associated with that point estimate; this is especially true for common benchmarks involving multiple tasks which are then merged into a single metric.

This is not to say that calculating the mean of the policy performance is erroneous; the rationale behind point estimating is that knowledge of the generic parameter θ (for example the mean) provides knowledge of the entire population which is described by the probability density or mass function $f(x, \theta)$ [14]. What we are missing, is knowledge of the population distribution.

Hypothesis test

According to [14], an hypothesis is *a statement about a population parameter*. As discussed in section 2.1.1, within the context of policy comparison, we are interested into n-samples tests with $n \geq 2$, i.e. testing whether two or more policies' performance differ significantly. This practice, which is considered standard in other fields of research, has not emerged yet into the machine learning and RL community.
The use of hypothesis testing emerged only recently in the RL literature following the so called "reproducibility crisis" [18]. An initial attempt was made by Henderson [32] suggesting the use of the standard t-test[61], a statistical test for comparing the mean of two samples under the assumption of equal variance and normality, and the Kolmogorov-Smirnov test, a test for comparing two samples and assessing whether they derive from the same distribution. This was later extended by [18] with the introduction of the Welch test.
Both the t-test and the Welch test assume the data to be normally distributed; this is often not the case which may lead to higher error rates within both tests, especially for lower sample sizes. Additionally, outliers skew the test statistics and result in a less powerful test [22].
Even the Kolmogorov-Smirnov test, which may alert the experimenter for possible non-normalities within the data, has a small power for low sample size [22].

The presence of outliers can be alleviated by trimming out the tails of the sample data; this approach has been proposed lately through the use of Inter-Quartile Mean [2]. From a hypothesis test point of view this translates into the Yuen-Welch test [75]: the test is a variant of the Welch test with the introduction of a trimming parameter which removes the sample tails.

We will briefly discuss the two most commonly used statistical tests in the following paragraphs. In doing so we will follow the lines of [48].

T test Let $X_1, \dots X_n$ and $Y_1, \dots Y_m$ be two normally distributed i.i.d. random variables with the same (unknown) variance σ^2 estimated by S^2. Let the parameter μ_x and μ_y be the expected value of the random variables which are estimated by $\bar{X}$ and $\bar{Y}$. The 2-sample t-test or Student's t test, formally tests the following null hypothesis:

$$H_0 : \ \mu_x = \mu_y \tag{2.5}$$

against the alternative hypothesis

$$H_1 : \ \mu_x \neq \mu_y \tag{2.6}$$

Let us define the test statistic t_{n+m-2} as follows:

$$t_{n+m-2} = \frac{\bar{X} - \bar{Y}}{S\sqrt{1/n + 1/m}} \tag{2.7}$$

Under the null hypothesis the test statistic 2.7 is distributed following Student's t distribution with $n + m - 2$ degrees of freedom.

Welch test The Welch test is a modified version of the t test which considers the case when $X_1, \dots X_n$ and $Y_1, \dots Y_m$ are two normally distributed i.i.d. random variables with the *different* (unknown) variances σ_x^2 and σ_y^2. The two variances are estimated by the parameter s_x and s_y respectively. Similarly to the t-test the null hypothesis is $H_0 : \ \mu_x = \mu_y$ and the alternative hypothesis $H_1 : \ \mu_x \neq \mu_y$.

The test statistics is defined as t_w [21]:

$$t_w = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \tag{2.8}$$

which again follows a t distribution with degrees of freedom defined as follows:

$$df_{welch} = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{s_x^4}{n^2(n-1)} + \frac{s_y^4}{m^2(m-1)}} \tag{2.9}$$

Bootstrap Confidence Intervals

We now turn to bootstrap confidence intervals: their use has been proposed in the RL literature by Colas et al. [18] to perform statistical comparison of policy end-scores. Bootstrap [24] is a statistical tool which draws multiple times a sample, with replacement, from the available data and uses this new samples to assess statistical accuracy [31].

The advantage of bootstrap CIs with respect to hypothesis tests, discussed in the previous section, is that we can compare bootstrap CI for the mean

without assuming any distribution. Additionally, the use of CIs, as opposed to p-values, results more intuitive.

Clearly bootstrap CI do aggregate data from different tasks; in addition as data is sampled with replacement there is no guarantee that all tasks are part of the bootstrap sample.
This criticality has been addressed lately by Agarwal et al [2] who recommended the use stratified bootstrap confidence intervals in order to estimate uncertainty in aggregate performance. The proposed stratified bootstrap approach is built as follows: we sample each task c times with replacement and aggregate this data. The resulting sample size is therefore equal to the number of runs c, times the number of tasks n. Note that this procedure aggregates data from different tasks however, with respect to classical bootstrap, each "strata" (i.e. each task) is sampled equally enforcing the bootstrap sample to contain data from all tasks.

2.1.4 Common Pitfalls and Challenges

The aim of this section is presenting the common pitfalls and challenges faced by the RL community, specifically with a statistical focus.

Small sample size

One major challenge in RL policy comparison is the use of small samples; in facts, due to the extensive computational requirements involved in benchmarking RL policies, recent literature have often compared policies based on just a handful of simulation runs [50, 49, 54]. The low sample size together with end-scores high variability have lead the community to conclusions which are not statistically sound and difficult, if not impossible, to reproduce [18].
Given the low sample regime, statistically informed decisions are even more important in order to make reproducible progress in the field.

Complex data distribution

When running a policy on a specific task, we may expect the resulting end-scores to be following a specific distribution. Surprisingly this is not always the case: simply running the same policy with different initial random seed can lead to learning curves apparently coming from different distributions [18].
The distribution of end-score rewards are affected by the random seed but clearly also by the policy and task themself: different policies result in different distributions of rewards and similarly, different reward distributions result when comparing the same policy on different tasks [15, 16]. An example is shown in appendix A which plots the distribution of the Atari 100k data reported in section 2.5: we clearly see a high variability of distributions both by changing policy or by changing task. Note that the best-fit distribution is mentioned in the plot.

The diversity of sample distributions poses challenges for RL policy comparison as most statistical models make some sort of assumption on the underlying distribution: for example the t-test, the Welch test and the Yuen test all assume normally distributed data.

End Score variability

RL tend to show high variability in end-score rewards [15, 16, 43, 27]: this is due to a range of factors such as the previously discussed random seeds [32], task or implementation detail. The variability has been reported to scale with performance [15]. Two common pitfalls which result in increased variability are:

- The use of biased metrics: as mentioned in section 2.1.3, some metrics in use are biased estimates of policy performance and contribute to the overall variability of RL end-scores [36]. Agarwal et al. [2] noted that using the sample mean is highly influenced by the performance of a few outliers, while sample median has high variability and is considered not a reliable metric [15].
  To overcome these issues the use of custom metrics have been proposed which target dispersion and risk or the use of ranks when comparing across tasks [15]. Others have proposed the use of robust metrics such as the interquartile mean (IQM) [2].

- Aggregating results from different tasks is an important source of variability: it was previously discussed how results from different tasks result are distributed differently [15, 16, 18, 11]. Indeed each task may have a different reward which results in completely different end-scores. In addition, tasks from the same benchmark (e.g. Atari) are often designed to provide a diverse set of tasks[12].
  This problem has been bypassed in literature through the use of normalization [12]: each end-score is transformed into a normalized score with the aim of comparing normalized scores across tasks. For the Atari benchmark the use of human records on each game and end-scores associated with a random policy are commonly adopted for scaling (meaning that a policy with normalized score equal to 0 behaves equally well as a random policy, whereas a normalized score higher than 1 would beat a human record). Whether the resulting normalized scores actually should be aggregated remains however dubious.
  A similar conclusion has been reported within the machine learning community: aggregating data from different data sets (within the context of RL, across different tasks/benchmarks) may lack of meaning when the tasks are not related. [22].

2.2 Statistical Blocking

The typical statistical experiment is concerned with classification of the data (hypothesis testing, point or interval estimates); classification is performed in

order to distinguish between two or more populations. Within the statistical literature we often speak of a treatment which refers to different levels of a particular treatment and generally is the variable that is under research.

Another type of categorization is however possible, which is called a block (or blocking factor). Blocks are categories the data is divided into but which are of no interest; the reason for dividing into blocks lies in that blocks provide an additional variability which if isolated can lead to more powerful statistical inference. Typically blocks were used in agriculture in order to control experimental variance induced for example when growing on different portions (blocks) of land [14].
Note that, although the terms *treatment* and *block* are standard within the field of statistics, this notation is not used through this work in order to better align with the RL literature. Instead, we will use the term policy to refer to the treatment and the term task or game to refer to the block.

We propose the application of blocking within the RL literature, an application which, to the author's knowledge, has never been proposed. The aim is to remove variability associated with comparing across multiple tasks or games; this consideration naturally leads to the use of blocking. Each task/game is thus considered a different block.

Blocking is typically used in the setting of ANOVA (which stands for Analysis of Variance and was first introduced by R. Fisher [26] to estimate means of several populations under normality assumption) and specifically in what is called a "Randomizes Complete Block Designs". Unfortunately, ANOVA is based on the assumption of normality of data and of sphericity (a property which requires the treatments to have the same variance). Both hypotheses are not met in the general RL framework: although it has been shown that the former may not affect the test dramatically, the latter gave greater effect and cannot be neglected [22].

Remark. The use of blocks "separates" the scores of individual tasks and in this way removes the need to normalize data before the analysis. The need to perform normalization was explained in section 2.1.4; in facts, when comparing scores across normalized tasks, we effectively introduce a arbitrary scaling factor which may have an impact on the resulting comparison. Take for example the Atari benchmark: it is customary to scale the results based on human scores on each task. However this approach may favor those tasks which are easier for the human player or, alternatively, specific "abilities".
From this perspective it is clear that the end-score normalization approach may introduce a bias in the comparison.

2.3 Skillings-Mack Test

This section introduces a novel procedure for RL policy comparison based on a statistical hypothesis test, a concept which was set forth in previous sections.

Before proceeding let us argue on the concept of statistical significance (thus on p-values); the use of p-values has been criticized both in the statistical community and within RL. From the RL literature, two major critics are set forth [2]:

1. The dichotomous nature (i.e. significant versus not significant). This is widely recognized as a distortion: clearly a p-value equal to 0.051 (not significant at the significance level of 0.05) is, in fact, identical to a p-value equal to 0.05 (significant at the significance level of 0.05).

   On the other hand there needs to be a threshold and a decision needs to be taken about whether policy π_1 performs better or not than policy π_2. This obvious fact should explicitly stated when performing research.

2. Common misinterpretations such as that lack of statistical significance does not demonstrate absence of effect or that given enough data any effect can be statistically significant. Later section 2.4 will describe a procedure which tackles this point and supports the use of hypothesis testing through the use of confidence intervals

To conclude, we believe these drawbacks are not limiting and instead we support the use of hypothesis testing for reliable and reproducible research.

We now introduces a generalization of the Friedman statistical test [28]; the resulting procedure has, to the author's knowledge, never been attempted within the RL literature.
The following treatment is based on the use of Skillings-Mack test [45, 60] for RL policy comparison. The methodology proposed by Mack and Skillings is an extension of Friedman's test to the case of incomplete blocks and higher replication within each block. The Skillings-Mack test can be shown to be equivalent to Friedman's test in the special case of a complete block setting [33]. The Skillings-Mack test was designed for statistical analysis involving two components and specifically for the relative location effect of one component within the various levels of the second component.

The test allows blocking and is thus ideally suited for multiple policy comparison on a set of tasks.

With respect to the RL framework, the Skillings-Mack test allows to compare different policies (1st component) within different tasks or games (2nd component), avoiding across task data aggregation. Through the treatment, reference will be made to the 26 tasks of the ATARI 100K benchmark, although the discussion holds in general for any number of tasks (i.e. blocks).

Remark. A note on notation. In literature it is common to refer to a *treatment* when discussing the first component and to a *block* when referring to the second component. This notation is not used through this work in order to better align with the RL literature. Instead, we will use the term policy to refer to the treatment and the term task or game to refer to the block.

2.3.1 Assumptions

This section discusses the assumptions which hold throughout this chapter. We will use the notation set forth in section 2.1.1. The following assumptions are enforced by the Skillings-Mack test procedure [60]:

1. Given a task i and a policy j, the random variables $X_{ij1} \ \ldots, \ X_{ijc_{ij}}$ are *i.i.d.* random variables with distribution function F_{ij}.

2. For each task i and policy j the data distribution share a similar structure given by the following equation:

$$X_{ij} = \mu + \beta_i + \tau_j + E \tag{2.10}$$

   where μ is the overall mean, the E are *i.i.d.* errors with distribution function F, β_i is the i-th block effect and τ_j is the j-th treatment effect. In other words, for different treatment-block combinations, the data may have arbitrary distribution but the effect of different treatment and blocks is additive. This may be represented as follows [33]:

$$F_{ij}(u) = F(u - \beta_i - \tau_j), \qquad -\infty < u < \infty \tag{2.11}$$

Note the similarity with ANOVA [26]: equation 2.10 is the similar to the one used in ANOVA, with the exception that there are no interaction terms and that the error term E_{ij} are not assumed normally distributed [45]. In other words, the procedure does not assume normally distributed data which makes this test particularly suited to analyze RL data.

Remark. Although the assumptions fit a wide range of problems, strictly speaking it is not possible to enforce assumption 2 in general as was discussed in section 2.5.2.

The procedure will be derived for the case of $c_{ij} > 1$; when the number of observations for each policy i and task j is equal to one, i.e. $c_{ij} = c = 1 \quad \forall i \in [1, \ldots n] \quad \forall j \in [1, \ldots, k]$, the Skillings-Mack test falls back to the classical Friedman test. We refer the reader to [33] for the case of incomplete blocks (i.e. missing data).

2.3.2 Test Procedure

The standard Skillings-Mack test involves testing the following null hypothesis:

$$H_0 : \ \tau_1 = \cdots = \tau_k \tag{2.12}$$

To this end we rank each policy within each task and average the resulting ranks across tasks. The null hypothesis 2.12 is rejected if a significative difference in ranks exists. For the simple case of a two-way comparison, this concludes the procedure having proved the two policies produce statistically significantly different results; for the scenario of $k > 2$ a further step is required to compare each policy pair. This is outlined in the following procedure, derived from [33, 45]:

1. For each task, rank each data point across policies; we represent by r_{ijl} the rank of point X_{ijl} with respect to all available data for task i, i.e. within the set $\left[X_{ijl} \quad | \quad i = \text{constant}, \quad j \in \left[1, \ldots, k\right] \quad l \in \left[1, \ldots, c_{ij}\right]\right]$. In case a tie occurs, average ranks are assigned. Table 2.2 visually shows the ranking of data.

Tasks	Policies			
	Policy 1	Policy 2	...	Policy k
Task 1	r_{111} ... $r_{11c_{11}}$	r_{121} ... $r_{12c_{12}}$	...	r_{1k1} ... $r_{1kc_{1k}}$
...	...	...	...	...
Task n	r_{n11} ... $r_{n1c_{n1}}$	r_{n22} ... $r_{n2c_{n2}}$	...	r_{nk1} ... $r_{nkc_{nk}}$

TABLE 2.2: Data rank

2. Let $q_i = \sum_{j=1}^{k} c_{ij}$ be the total number of observations within block i and let $ra_{ij} = \sum_{l=1}^{c_{ij}} r_{ijl}/q_i$ be the cell-wise weighted averaged of ranks, as shown in table 2.3: we obtain a possibly non-integer rank for each policy on each task. For each policy j, compute the sum of ranks across tasks:

$$S_j = \sum_{i=1}^{n} ra_{ij} = \sum_{i=1}^{n} \sum_{l=1}^{c_{ij}} \frac{r_{ijl}}{q_i} \tag{2.13}$$

Task	Policies			
	Policy 1	Policy 2	...	Policy k
Task 1	ra_{11}	ra_{12}	...	ra_{1k}
...	...	...	...	...
Task n	ra_{n1}	ra_{n2}	...	ra_{nk}

TABLE 2.3: Average Policy rank per task

3. Let us define the vector $\mathbf{S}$ as the sums of cell-wise average ranks centered about their expected value under the null hypothesis:

$$\mathbf{S} = \left[S_1 - E_0\left(S_1\right), \ldots, S_k - E_0\left(S_k\right)\right] \tag{2.14}$$

where $E_0(\cdot)$ is the expected value under the null hypothesis and is equal to:

$$E_0\left(S_j\right) = \sum_{i=1}^{n} \frac{c_{ij}\left(q_i + 1\right)}{2q_i} \tag{2.15}$$

Because the S_j are linearly dependent, we could omit one. We will, without loss of generality, omit S_k. In doing so, we obtain the linearly independent vector $\tilde{\mathbf{S}}$.

4. The covariance of vector $\tilde{\mathbf{S}}$ under the null hypothesis has the form:

$$\mathbf{\Sigma}_{\tilde{\mathbf{S}},0} = [\sigma_{s,t}] \tag{2.16}$$

where

$$\sigma_{s,t} = \begin{cases} \sum_{i=1}^{n} \frac{c_{is}(q_i - c_{is})(q_i+1)}{12q_i^2} & s = t = 1, \ldots, k-1 \\ \sum_{i=1}^{n} \frac{c_{is}c_{it}(q_i+1)}{12q_i^2} & s \neq t, \quad s,t = 1, \ldots, k-1 \end{cases} \tag{2.17}$$

Let $\mathbf{\Sigma}_{\tilde{\mathbf{S}},0}^{-1}$ be its inverse, then the Skillings-Mack test statistic is defined as:

$$MS = \tilde{\mathbf{S}}\mathbf{\Sigma}_{\tilde{\mathbf{S}},0}^{-1}\tilde{\mathbf{S}}' \tag{2.18}$$

This is greatly simplified in the case of $c_{ij} = c$ as follows:

$$MS = \left(\frac{12}{k(C+n)}\right) \sum_{j=1}^{k} \left(S_j - \frac{C+n}{2}\right)^2 \tag{2.19}$$

where

$$C = \sum_{i=1}^{n}\sum_{j=1}^{k} c = cnk \tag{2.20}$$

and

$$S_j = \sum_{i=1}^{n} ra_{ij} = \sum_{i=1}^{n}\sum_{l=1}^{c} r_{ijl} \tag{2.21}$$

5. The null hypothesis 2.12 is rejected if

$$MS \geq ms_\alpha \tag{2.22}$$

where the constant ms_α is computed to allow the type-I error probability equal to α. The computation of the critical value is deferred to section 2.3.3.

6. For the case $c_{ij} = c$ and $k > 2$, if the null hypothesis 2.12 is rejected, an additional step is required in order to compare pairs of policies and define which are significantly different. Consider two policies j and j'. Decide that $\tau_j \neq \tau_{j'}$ if:

$$\left|S_j - S_{j'}\right| \geq \left(\frac{k(C+n)}{12}\right)^{\frac{1}{2}} q_\alpha \tag{2.23}$$

where q_α is the upper αth percentile for the distribution of the range of k independent variables normally distributed with mean 0 and variance 1. The right hand side of equation 2.23 will be called critical difference or CD following the convention adapted by [22] for the Nemenyi Test [51].

The Skillings-Mack test is suited to cover applications with few data: the original paper demonstrated that the procedure is more efficient [25] than ANOVA's F test in many cases and nearly as efficient even when the sampling distribution is normal [45]. Comparison of the method with other methodologies is performed in chapter 3.

2.3.3 Skillings-Mack critical value

This section discusses the computation of the Skillings-Mack test critical value. Three implementations are provided within the custom developed code in the Annex B: this implementation matches the one provided by Hollander in their R package NSM3 [34]. The following methods are implemented:

1. Exact Method;
2. Asymptotic method;
3. Monte Carlo approximation.

these are discussed in the following:

Exact Method This is used only for small sample size. The procedure is as follows:

1. Assuming that all possible rankings are equally probable under the null hypothesis, we build a dataset containing all possible permutations of the rankings; this is possible since the number of ranks is low.
2. For each permutation, the MS statistic 2.19 is calculated; In this way we obtain its exact distribution.
3. It is thus possible to calculate the critical value of the MS test statistic which corresponds to the 5% (significance level) in the right tail of the exact distribution.

As we are building all possible permutations of the rankings, this method rapidly becomes computationally infeasible as the number of ranks increases.

Asymptotic Method According to [45] for large samples, the MS test statistic is χ^2 distributed with $k-1$ degrees of freedom. This allows rapid computation of the critical value; additionally it has been noted that the asymptotic method is adequate even at significance level 0.05 and with small sample sizes ($c_{ij} > 3$). At lower significance levels the asymptotic approximation becomes more conservative unless the sample sizes are large [33].

Monte Carlo approximation The Monte Carlo method is an approximation of the Exact method discussed earlier. The procedure is similar, with the only difference that not all rank permutations are calculated; instead we randomly generate a permutation M times with M a user defined number (defaults to 10000 in the implementation). The method then proceeds with the computation of the MS test statistic and its distribution similarly as done in the Exact method.

2.4 Inferential Confidence Intervals

Inferential confidence intervals (ICIs) have been proposed by Goldstein and Haley [29] with the aim of testing statistical difference, equivalence and indeterminacy in a way which is equivalent to a standard null hypothesis statistical test [67]. The approach was suggested in order to avoid some of the misuses of null hypothesis statistical testing: ICI are based on confidence intervals which are easily represented and interpreted.

The use of ICI is straightforward: if two inferential confidence intervals do not overlap, then a statistically significant difference can be reported. Using 95% confidence intervals[1] (CI) and drawing a conclusion $p < 0.05$ for the related test is not guaranteed in general so we correct the CI by a factor ϵ. In general we may represent this as follows:

$$CI = [t - \gamma_{low},\ t + \gamma_{up}] \qquad \longrightarrow \qquad ICI = [t - \epsilon\gamma_{low},\ t + \epsilon\gamma_{up}] \tag{2.24}$$

where t is an estimate of a population parameter (e.g. the mean) and $\gamma_{up}, \gamma_{low}$ define the the upper and lower bounds of the CI. The calculation of factor ϵ depends on the underlying test which is assumed; from this perspective, using ICIs is the same as using the corresponding test which has been used to calculate the correction factor ϵ. It is therefore essential to specify which test is being used; formally the definition of ICI holds when the assumptions of the underlying test also hold.

The use of ICI was initially based on the t-test [29, 67] and later extended to the Welch test.
Let us assume a hypothesis test for the difference of the means with $t_2 > t_1$. We depart from the work of Marasini et al. [46] who applied a generalization [68] of the framework for calculating the correction factor ϵ to bootstrap CI as follows (section 3.3 of [46]):

$$\left[t_2 - \epsilon\left(t_2 - t_{2\,low}\right)\right] - \left[t_1 + \epsilon\left(t_{1\,up} - t_1\right)\right] = d \tag{2.25}$$

this equation holds when the parameter $t_2 > t_1$. It is possible to recognize that the left hand side of the equation represents the difference between the extremes of the two CIs. The equation forces the gap between the two CIs to

[1] Following the convention of [68] we will call these confidence interval "descriptive" to distinguish them from the "inferential" confidence intervals introduced.

be equal to d; given this parameter the correction factor becomes:

$$\epsilon = \frac{t_2 - t_1 - d}{t_2 - t_1 + t_{1\ up} - t_{2\ low}} \tag{2.26}$$

where ϵ is the correction factor to be applied to equation 2.24, t_1 and t_2 are the central values of the two confidence interval, $t_{1\ up}$ and $t_{2\ low}$ are respectively the upper and lower bound of the confidence interval for t_1 and t_2.
Within [46] the parameter d is selected to be $P - \alpha$, where P is the level α p-value for the selected null hypothesis which results in the following:

$$\epsilon = \frac{t_2 - t_1 + P - \alpha}{t_2 - t_1 + t_{1\ up} - t_{2\ low}} \tag{2.27}$$

Equation 2.27 can be used together with the Skillings-Mack test procedure developed in section 2.3 under the assumption of asymptotic behaviour. We will not use standard t-test or Welch test as their assumptions are often violated. The assumption of asymptotic behaviour is needed to compute the Skillings-Mack probability distribution in order to calculate p-values: from [33] we know that the statistic is χ^2 distributed, with $k - 1$ degrees of freedom: we can thus compute the value $P - \alpha$.

Modified approach Through this work we will use a slightly modified approach to ICIs. This modification is introduces to tackle the behaviour of equation 2.27 under specific circumstances: when the value of $t_2 - t_1 \leq \alpha$ then the numerator of the equation may go to zero or get negative for low p-values. In this scenario, the equation 2.27 would result in a negative value of ϵ which is indeed not correct as it would suggest to flip the lower confidence bound over the mean of the bound, making it in facts the upper confidence bound.
Additionally, the imposed distance between confidence interval extremes is arbitrary when the null hypothesis is rejected and not descriptive of any real effect. This is related to how ICIs are built: ICIs require the CI to be non-overlapping when the underlying test null hypothesis is rejected but do not specify the amount of overlap. For this reason, we prefer to return to descriptive confidence intervals when possible instead of arbitrarily fixing the distance between CI extremes.

We introduce the following modifications:

1. We add a scaling factor to the value of $d = P - \alpha$. In order to keep the comparison with $t_2 - t_1$ we scale d by the same mean difference $t_2 - t_1$.

Equation 2.27 then becomes:

$$\begin{aligned}\epsilon =& \frac{t_2 - t_1 + (t_2 - t_1)(P - \alpha)}{t_2 - t_1 + t_{1\ up} - t_{2\ low}} = \\ =& \frac{(t_2 - t_1)(1 + P - \alpha)}{t_2 - t_1 + t_{1\ up} - t_{2\ low}} = \\ =& \frac{(t_2 - t_1)(1 + P - \alpha)}{(t_2 - t_1)\left(1 + \frac{t_{1\ up} - t_{2\ low}}{t_2 - t_1}\right)} = \\ =& \frac{1 + P - \alpha}{1 + \frac{t_{1\ up} - t_{2\ low}}{t_2 - t_1}}\end{aligned} \tag{2.28}$$

The need to scale this parameter was indeed presented by Marasini et al. [46].

2. Whenever the test is not statistically significant or when the two confidence intervals are non-overlapping we keep the descriptive confidence interval, that is we avoid computation of ϵ and keep its value to 1.

CIs can be obtained through bootstrap [18] or stratified bootstrap [2].

2.5 Application to Reinforcement Learning

The next sections will demonstrate the application of the proposed methodologies to a specific RL example, selected to represent the state of the art policy comparison approach. Generalization of these results are left to chapter 3.

Before proceeding we discuss and analyze the example's data. In order to facilitate comparison with current state of the art results, this section will make use of an available dataset which was made available by Rishabh Agarwal [1]; within his work [2], he proposed the use of stratified bootstrap confidence intervals to compare policies and applied it to the Atari 100k benchmark. Six different policies were trained on the 26 games of the benchmark resulting in 100 data points (final scores/rewards) for each policy-game combination, a total of 15600 simulation points. Out of this dataset, we use the following:

- 5 different RL policies, namely: CURL [41], DER [69], DrQ(ϵ)[40, 2], OTR [30], SPR [55];
- 26 different tasks or games, namely: Alien, Amidar, Assault, Asterix, BankHeist, BattleZone, Boxing, Breakout, ChopperCommand, CrazyClimber, DemonAttack, Freeway, Frostbite, Gopher, Hero, Jamesbond, Kangaroo, Krull, KungFuMaster, MsPacman, Pong, PrivateEye, Qbert, RoadRunner, Seaquest and UpNDown [39].

A summary of the available data is shown in 2.4 highlighting the availability of 2600 data points for each policy and a total of 13000 total simulation points.

Tasks	Policies				
	CURL	DER 2	DrQ(ϵ)	OTR	SPR
Alien	100	100	100	100	100
Amidar	100	100	100	100	100
...	...	...	...	...	...
UpNDown	100	100	100	100	100

TABLE 2.4: Number data points available

2.5.1 Data split

When performing statistical tests, we often require data to be independent. The use non-independent data is a known pitfall in machine learning literature, commonly avoided by using k-fold cross validation [52]. We take a similar approach throughout this work allowing the 100 data points available for each policy and task pair to be divided into separate samples, with sample size of 3, 5, 10 25 and 50 respectively. In other words, for each policy-task combination we sample the corresponding data without replacement in order to form 5 different datasets with respectively 3, 5, 10, 25 and 50 datapoints for each policy-game combination. The resulting datasets contain 468, 780, 1560, 3900 and 7800 simulation points respectively.

2.5.2 Data Distributions

In this section we provide insight into the available data and demonstrate the nature of the challenges discussed in section 2.1.4. We begin with evaluating normality and then we show the distribution of data.

Normality

As was highlighted in section 2.1.4, RL end-scores generally follow complex non-normal distributions. We question this assumption and check whether data is normally distributed. As the number of normality tests equals to 156 (i.e. 26 games for 6 policies) plotting Q-Q graphs or histograms looks not feasible. Instead we test the null hypothesis that the data comes from a normal distribution, based on the D'Agostino and Pearson's test [20]. The result is shown in Figure 10 where a color code is added based on the resulting p-value which is annotated on top.

Assuming $\alpha = 0.05$, Figure 10 shows that the null hypothesis cannot be rejected in 49 samples out of 156; only roughly 1 distribution out of 3 can be assumed normal. Two further considerations can be drawn:

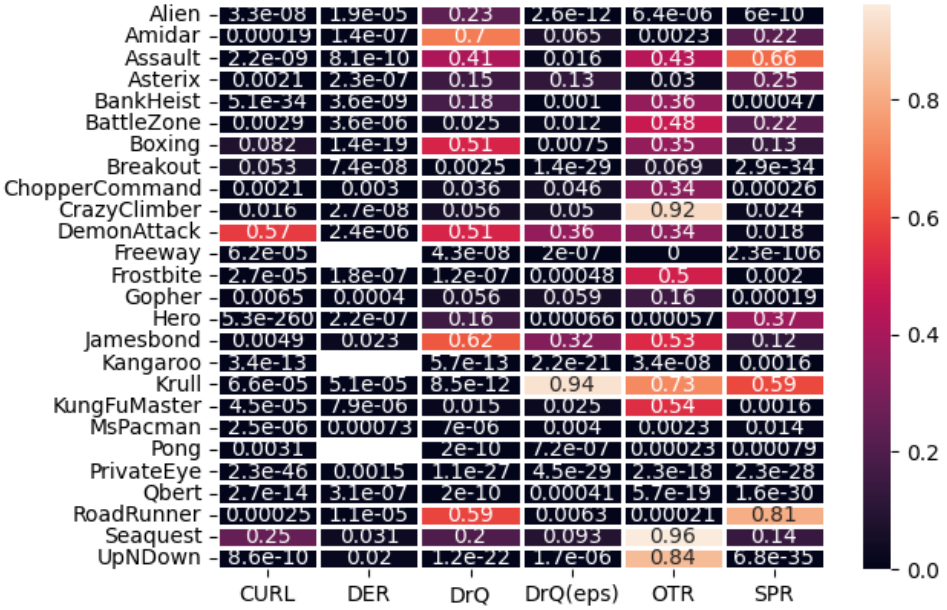


FIGURE 10: Test null hypothesis that data are normally distributed

1. It is not possible to highlight a task whose reward scores are normally distributed across all policies. In fact, normality of reward scores seems to be more correlated with the policy rather than the task.

2. Highly non-normal distributions are common. For instance, 70 out of 156 samples have a p-values less than 0.001.

Best fit distribution

Given the results of previous section, the obvious question arises about how data is distributed. To do so, we use the python *distfit* package (version 1.7.1) and allow the fit for popular distributions. When the normality test fails, we allow the algorithm to select the best fit distribution; for non-significant p-values we force the use of the normal distribution. The result is shown in Appendix A - Atari 100k distributions.

2.5.3 Reference scores

In order to compare the results of our proposed methodology, this section presents available state of the art policy comparison methods; the use of stratified bootstrap confidence intervals (CIs) on the data presented in section 2.5 has been demonstrated [2]. This result is reported in Figure 11 for comparison. The plot shows calculated median and interquartile mean for 5 policies (represented with different colors) respectively on 3, 5, 10, 50 and 100 samples (runs). The data mimics what was discussed in section 2.5.1 with the only difference that the datasets are not independent (e.g. the 100 run point contains all data from the other runs).

2.6 Skillings-Mack Test application

In this section we apply the Skillings-Mack test procedure outlined in section 2.3. We run the procedure multiple times with increasing $c_{ij} = c$, each with a separate sample as discussed in section 2.5.1. We start by applying the procedure to a 2-policy scenario first and then to a complete scenario with all five policies.

We start with the two policy comparison: we will use policies SPR and DrQ(ϵ) for this comparison but the same procedure applies in general. It shall be noted that this case leads back to the null hypothesis 2.1 and the alternative hypothesis 2.3, which provide clear evidence of whether the difference between the two policies is statistically significant: point 6 from section 2.3.2 shall therefore not be applied.

We apply the steps from section 2.3.2: the policy sum of rankings S_j according equation 2.21 is computed; follows the computation of the test statistic MS according to equation 2.19. This results in the MS score shown in Table 2.6.

Number of runs	Sample size	MS scores
3	78	5.0
5	130	15.1
10	260	21.2
25	650	61.5
50	1300	96.2

TABLE 2.5: Skillings-Mack test statistic for the SPR-DrQ(ϵ) policy comparison.

The critical value at $\alpha = 0.05$ is computed using the χ^2 distribution with 1 degree of freedom (refer to section 2.3.3, asymptotic method)[2]:

$$ms_\alpha = 3.8 \tag{2.29}$$

Clearly, the computed MS score is much greater than the required ms_α for all values of c so it is possible to reject the null hypothesis. We conclude that the two policies are significantly different.

We now turn to the comparison between all the 5 policies together. Although this approach allows us to compare with results from other papers, we note that the procedure of section 2.3 is overly conservative when comparing one novel policy against multiple state of the art policies; in facts, the methodology adjusts the critical value based on the number of comparisons which are $k(k-1)/2$ whereas when comparing against one "novel" policy

[2]The critical value as computed with the Monte Carlo method differs only by 1.7% with sample size equal to three. This difference does not modify the conclusions; we will therefore keep the asymptotic calculation for ease of replication.

we only need to run k comparisons. Reducing the number of comparisons is preferred, because when performing multiple comparisons one has to account for the added error probability of running multiple tests (e.g. through the Bonferroni correction [13]).

The procedure is the same outlined for the 2-sample case. This resulting calculated MS statistics are shown in Table 2.6.

Number of runs	Sample size	MS
3	78	113.1
5	130	180.6
10	260	323.3
25	650	956.5
50	1300	1866.0

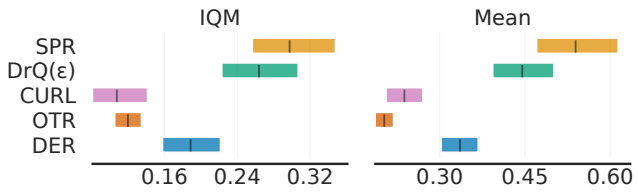
TABLE 2.6: Skillings-Mack test statistic for the comparison of all 6 policies

The corresponding critical value at $\alpha = 0.05$ is calculated through the χ^2 with 4 degrees of freedom distribution as follows:

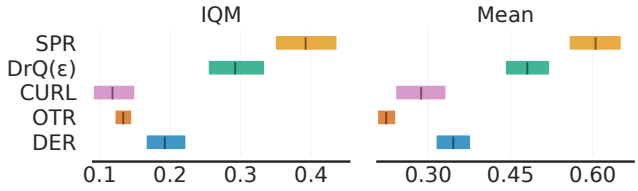
$$ms_\alpha = 11.1 \tag{2.30}$$

Again, the computed MS score is largely greater than the required ms_α for all values of c so it is possible to reject the null hypothesis.

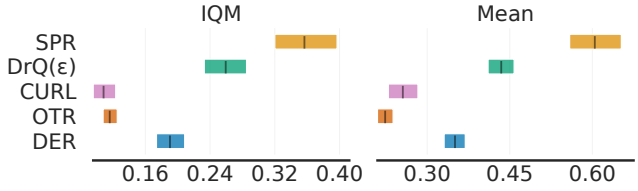
To proceed with the pairwise comparison, we follow the approach outlined in section 2.3.2 and plot a ranking plot with a critical difference CD calculated according equation 2.23. This is shown in Figure 12. Note the plots show lower numbers on the right and assumes ranking 1 to be the best value a policy can obtain.



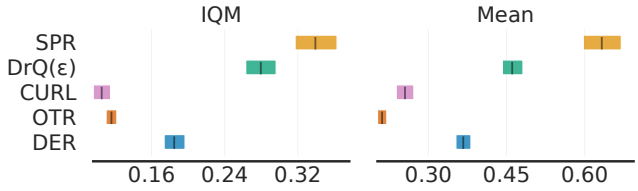
(A) 3 replications



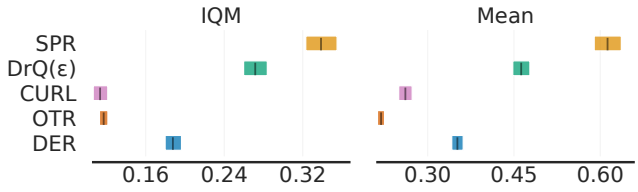
(B) 5 replications



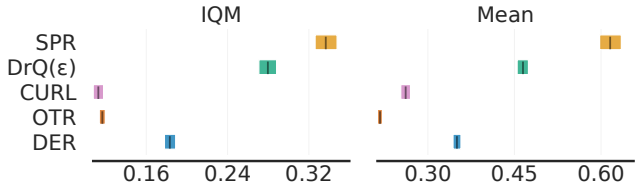
(C) 10 replications



(D) 25 replications



(E) 50 replications



(F) 100 replications

FIGURE 11: CIs for increasing sample sizes (replication runs), following [2]. Abscissa reports end-scores

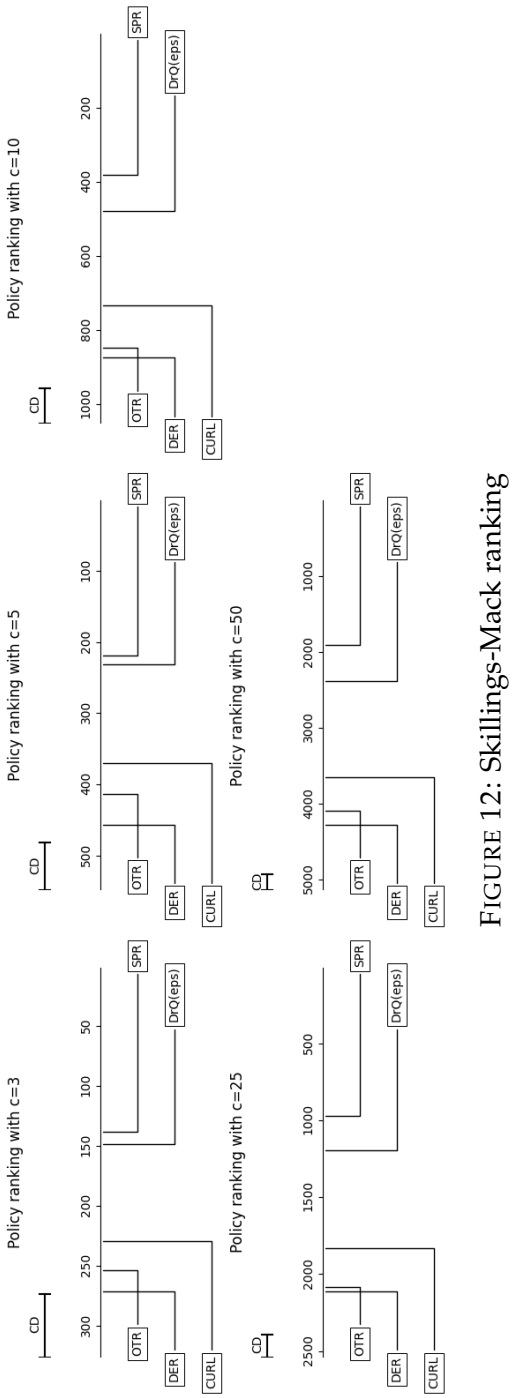


FIGURE 12: Skillings-Mack ranking

The methodology proposed leads to a series of advantages with respect to the state-of-the art methodologies applied in the RL literature:

- The procedure produces more statistically significative differences w.r.t. stratified bootstrap CI based on the interquartile mean (IQM): in facts, whereas stratified bootstrap CIs still overlap at $c = 10$ (see figure 11), this procedure produces statistically significant different results with the same number of data points. Comparing policy SPR with DrQ(ϵ) at $\alpha = 0.05$ we note from table 2.5 that the difference is statistically significative even when $c = 3$ [3]. When all five algorithms are tested together, we can infer a statistically significative difference at $c = 10$, where we note how the difference $\left|S_{SPR} - S_{DrQ(\epsilon)}\right| = 122.2$ is bigger than the critical difference $CD = 113.5$.

- Different task scores are not aggregated, resulting in a more interpretable and safe statistic, as was discussed in section 2.1.4. This is especially helpful when completely different benchmarks are compared which do not share common scores.

- The procedure does not need to normalize tasks. Because normalization is normally carried out based on some user-defined limits, this process introduces weights into the comparison, arbitrarily inflating or deflating the scores of one task against another. Although the use of human scores to normalize the data may seem reasonable, this might introduce a bias towards specific tasks, based on how well a human can handle it.

- By comparing the critical difference plotted on figure 12 with table 2.5 we note how the latter provides statistically significant results even at low sample size, whereas the former does not. This should be expected as the critical difference method actually performs a multiple comparison test and therefore requires stricter type-I error control.

With respect to the approaches based on the comparison of scores, the use of a ranking system does not provide direct evidence of the resulting score. Although this may be less interpretative it is believed that this approach does not lead to the confusion of assigning one single score to the policy which may not be meaningful when multiple tasks are aggregated. Our suggestion is to base information on score statistics computed for individual tasks only.

2.7 Stratified Bootstrap ICI application

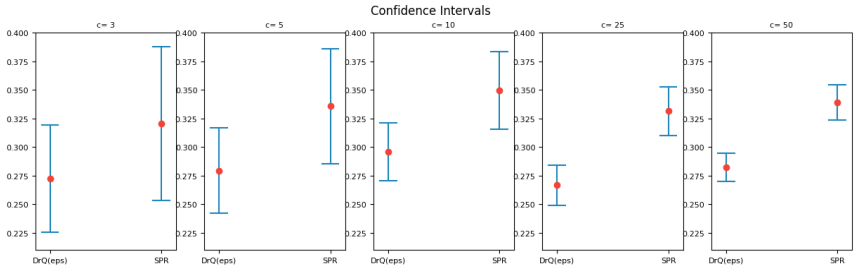
This section applies the concept of Inferential Confidence Intervals developed in section 2.4 to the data shown in section 2.5.

[3] Note that replication number $c = 3$ does not mean the sample has size 3; in order to compute the sample size we have to multiply the replication number c by the number of games available (i.e. 26 in this case). The resulting sample size is thus 78.

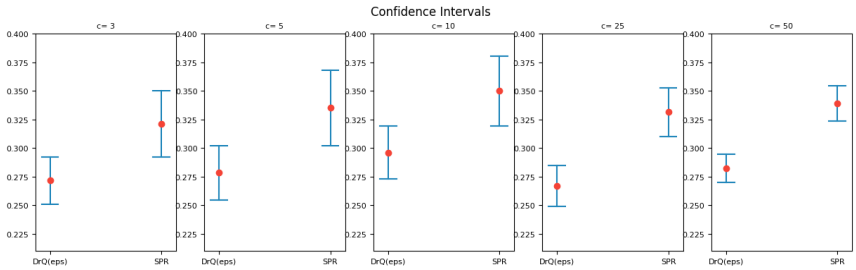
The ICI procedure is here applied to the stratified Bootstrap CIs [2] based on the sample interquartile mean. These CI have been proved to accurately represent policy performance and are considered state of the art; in addition we will be able to compare with results from section 2.6.
We firstly compute the correction factor ϵ according to equation 2.27. Assuming a χ^2 distribution for the MS statistic, we can obtain p-values for the computed MS values of table 2.5; This is shown in table 2.7 where we note all p-values are less than 0.05, given the fact that the critical value for MS is 3.84.

Number of runs	Sample size	MS scores	$\alpha - P$	ϵ
3	78	5.0	0.024	0.44
5	130	15.1	0.050	0.65
10	260	21.2	0.050	0.94
25	650	61.5	0.050	1.0
50	1300	96.2	0.050	1.0

TABLE 2.7: ICIs parameters for the skillings-mack test



(A) Descriptive CIs



(B) Inferential CIs

FIGURE 13: Stratified Bootstrap CIs and corresponding ICIs for policies $DrQ(\epsilon)$ and SPR calculated with [3]

The stratified bootstrap CIs are calculated based on a random sample from the dataset presented in 2.5 and are shown in figure 13a. The sample has been selected in order to demonstrate different overlapping conditions of the two CIs.

Following the procedure outlined in 2.4 we obtain the values for ϵ described in table 2.7. The resulting ICIs are shown in figure 13b.

Within the RL literature, the use of CI for inference is increasingly being suggested [18, 2] but an approach based on ICIs has never been proposed to the author's knowledge. ICIs have been developed with the aim of supporting *inference* and provide results matching with statistical hypothesis testing. The added benefit lies in the inference "by eye" which ICIs make possible: intuitively, if ICIs do not overlap then the difference between two policies can be said to be statistically significant.

With respect to stratified bootstrap results presented in figure 11, we note that the skillings-mack test provides statistically significant result even with 3 runs per game[4]; this leads the ICIs built on top of this test to provide non-overlapping CI for all sample sizes analyzed. By keeping the same CIs representation, the introduction of ICIs allows to perform inference "by eye" at a specific significant level.

[4]We remind that the sample size is not 3; the number of runs must be multiplied by the number of games within the benchmark, in this case 26, resulting in a total of a sample dimension of 78 for each policy.

Chapter 3

A Simulation Analysis of Statistical Methods for RL Policy Comparison

Chapter 2 discussed common methods for Reinforcement Learning (RL) policy comparison and proposed two novel methodologies based on inferential confidence intervals (ICIs) and statistical blocking paired with the Skillings-Mack non-parametric statistical test; these procedure were tested against an available benchmark in literature and showed more significative results over Stratified Bootstrap Confidence Intervals. The objective of this chapter is to make one step further and generalize the discussion.

3.1 Approach

In order to generalize the results obtained in the previous chapter, the RL policy comparison methods need to be tested with respect to synthetic datasets which results are known a-priori. Given such a dataset, there are two possibilities: either the null hypotheses 2.1, 2.2 hold or the alternative hypotheses 2.3, 2.4 hold. In the latter case, it is essential to specify how large the difference between the different policies is: larger differences are evidently easier to distinguish.

The generation of samples from a known distribution in order to comply with one of the two hypothesis is discussed in detail in section 3.2. By enforcing one of the two conditions it is possible to compare the expected outcome with the actual one and evaluate whether the method successfully predicted the selected hypothesis. Practically speaking, it is customary to control the following two parameters:

- The probability type-I error (i.e. of a false positives) which is the probability the test incorrectly rejected the null hypothesis or, in other words, the probability that, true the null hypothesis, the test gives statistically significant results.
- the statistical power (i.e. the probability of a true positive) which is the probability of correctly rejecting the null hypothesis or, in other words, the probability the data follows the alternative hypothesis and

the method correctly predicts this. This can be expressed as 1 minus the type-II error probability.

The estimates of the error probabilities is performed by running a simulation with a large number of replications; in the next sections, it will be assumed, unless otherwise stated, that a total of $N_s = 5000$ replications are performed. The resulting error probability is estimated by counting the relative number of occurrences for a specific case.
For type-I error we run the simulation under the null hypothesis and compute the percentage of replications which resulted into the null hypothesis being rejected; this percentage should match with the imposed α. For statistical power we enforce the alternative hypothesis (i.e. the means are effectively different) and count the percentage of simulations which resulted into the test rejecting the alternative hypothesis; this percentage should be as high as possible.

The estimate of type-I error probability and statistical power has been attempted within the RL policy comparison literature by Colas et al. [17]. Colas' contribution pointed out which RL policy comparison method performed best under the assumption of data being generated from a single distribution. This work, on the other hand, captures the complexity inherited from multi-task comparisons as discussed in section 2.1.

The approach followed through the remaining of this section is as follows:

- In section 3.2 we provide an overview of how data is being generated.
- Section 3.3 discusses the simple case of two policies with tasks following the same normal distribution.
- Section 3.4 covers the case where tasks follow the same distribution across policies; we call this case homogeneous because task distribution is the same (i.e. homogeneous) between policies. We allow tasks to have different distributions within the same policy and we do not assume normality.
- Within section 3.5, the general case is being considered where variability is introduces across policies either in the form of separate distributions or as different variance.

3.2 Synthetic data generation

This section discusses the generation of a synthetic data with the aim of building samples for RL policy comparison.
Within the literature, the comparison of RL policies was based either on a metrics on a single task or aggregated metrics over multiple tasks [17]; this data can be regarded as originating from a single population distribution. Colas et al [17] calculated type-I and type-II error probabilities under this

scenario for multiple distributions and different policy comparison methods, including the t-test, Welch's test, Wilcoxon Mann-Whitney rank sum test, Bootstrap confidence interval test and others.

Within this work an attempt is made to introduce differences induced by the multiplicity of different tasks and policies, each with its different distribution. We stress the importance of leaving open the possibility, for each task, to select a different distribution; this applies not only across tasks but also across policies (i.e. a different distribution can be selected for the same task but different policy as was recognized by Colas et al. [18]).

Formally, we ask to define each random variable X_{ij} as defined in section 2.1.1 by its distribution. Clearly it is still possible to lead back this approach to a standard single-distribution by enforcing the X_{ij} to follow the same distribution.

Each X_{ij} is defined by the distribution family and its corresponding parameters; a selection of common distributions has been used and is listed hereafter:

- Normal distribution, parameterized by variables location (loc) and scale (sc);
- Student's t distribution, parameterized by variables location (loc), scale (sc) and degrees of freedom (df);
- Exponential distribution, parameterized by variables location (loc) and scale (sc);
- Log-normal distribution, parameterized by variables location (loc), scale (sc) and the shape parameter(s);
- Gamma distribution, parameterized by variables location (loc), scale (sc) and the shape parameter (a);
- Beta distribution, parameterized by variables location (loc), scale (sc) and two shape parameters (a) and (b);
- Double Weibull distribution, parameterized by variables location (loc), scale (sc) and the shape parameter (c);
- Pareto distribution, parameterized by variables location (loc), scale (sc) and the shape parameter (b);

These distributions are modelled through the use of the *scipy* library [19].
In order to compute type-I error probabilities and the statistical power, we need to test the method under two specific cases:

1. When the null hypothesis 2.2 is true;
2. When the null hypothesis 2.2 is false.

To this end, each distribution needs to be translated and scaled so to allow the selected hypothesis to hold. For each random variable we therefore enforce the mean m and variance v. This is performed by modifying the location loc and scale sca parameters as follows:

- Normal distribution;

$$loc = m \tag{3.1}$$
$$sca = \sqrt{v} \tag{3.2}$$

- Student's t distribution;

$$loc = m \tag{3.3}$$
$$sca = \sqrt{\frac{v(df-1)}{df}} \tag{3.4}$$

- Exponential distribution;

$$loc = m - \sqrt{v} \tag{3.5}$$
$$sca = \sqrt{v} \tag{3.6}$$

- Log-normal distribution;

$$loc = m - e^{\frac{s^2}{2}} \sqrt{\frac{v\, e^{s^2}}{e^{s^2}-1}} \tag{3.7}$$
$$sca = \sqrt{\frac{v\, e^{s^2}}{e^{s^2}-1}} \tag{3.8}$$

- Gamma distribution;

$$loc = m - a\sqrt{\frac{v}{a}} \tag{3.9}$$
$$sca = \sqrt{\frac{v}{a}} \tag{3.10}$$

- Beta distribution;

$$loc = m - \frac{a}{a+b}\sqrt{\frac{v\,(a+b+1)(a+b)^2}{a\,b}} \tag{3.11}$$
$$sca = \sqrt{\frac{v\,(a+b+1)(a+b)^2}{a\,b}} \tag{3.12}$$

- Double Weibull distribution;

$$loc = m \tag{3.13}$$

$$sca = \sqrt{\frac{v}{\Gamma(1+\frac{2}{c})}} \tag{3.14}$$

where $\Gamma(\cdot)$ is the gamma function.

- Pareto distribution;

$$loc = m - \frac{b}{b-1}\sqrt{\frac{v(b-1)^2(b-2)}{b}} \tag{3.15}$$

$$sca = \sqrt{\frac{v(b-1)^2(b-2)}{b}} \tag{3.16}$$

We will then plot actual significance levels as discussed in section 3.1 at nominal significance level of $\alpha = 0.05$.

In order to test different conditions the following sections will plot results with increasing sample size. To align with the notation used in RL we use the number of runs for each task as the abscissa. In other words this is the number of simulations performed for each task-policy combination. This should not be confused with the total sample size which is obtained by multiplying the number of runs by the number of tasks used. To facilitate interpretation a secondary axis is added on the top of all plots showing the aggregate sample size.
Unless otherwise noted, we will assume the following single block-policy sample size (number of runs): 2, 3, 5, 10, 15, 20, 30, 40, 50, 65, 80, 100.

Through the following sections we will make use of the python package scipy [19] for performing the following tests:

- **t-test** through the function *stats.ttest_ind*.
- **Welch test** through the function *stats.ttest_ind* with parameter *equal variance* set to *False*.
- **Yuen test** through the function *stats.ttest_ind* with parameter *equal variance* set to *False* and the additional trimming parameter set to 0.1.

The implementation of stratified bootstrap confidence intervals has been taken from [3]; this implementation computes stratified bootstrap CIs of four statistics: the inter-quartile mean, the median, the mean and the optimality gap [2]. In the following we will report the inter-quartile mean, which is the suggested approach of the original paper and the mean.
The skillings-mack procedure used in the following has been developed by the author for the case $c_{ij} = c$. The Python code is available in the appendix.
The implementation of ICIs was discussed in section 2.4 and is available in

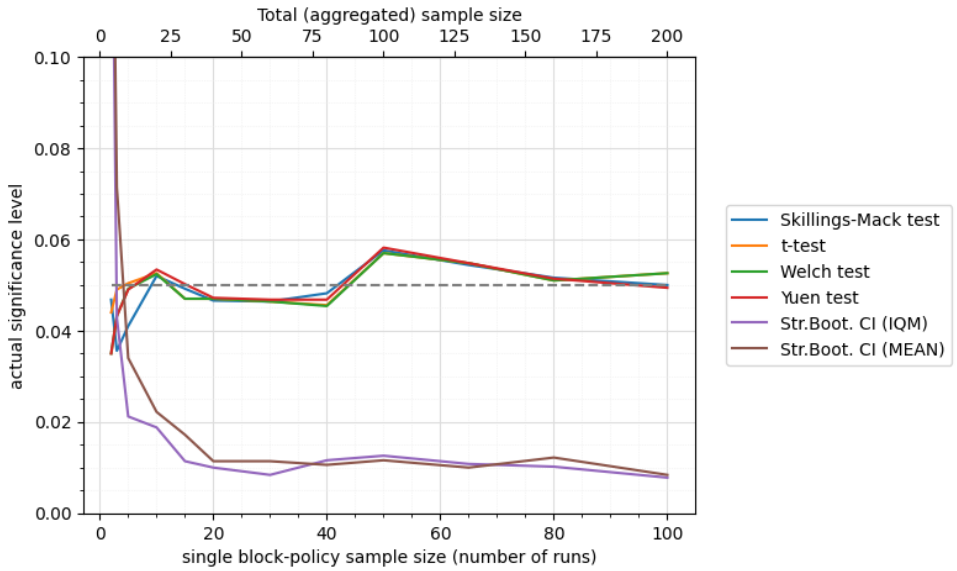


FIGURE 14: Actual significance level vs sample size of simulation with 5000 replications with two normally distributed tasks with same mean and variance. Nominal $\alpha = 0.05$.

the Python code within the appendix. We base the computation of ICI on the Skillings-Mack test. However, this will not be shown within the following sections in order to streamline the presentation of results: in facts, actual significance levels of ICIs based on the Skillings-Mack test always overlap with the latter and are therefore redundant.

3.3 Normally distributed tasks

The comparison between RL policy comparison methods starts with the simplest case of normally distributed data with same mean and variance. We start with $n = 2$ tasks and $k = 2$ policies as represented in table 3.1; this scenario is ideally the same assumed by the t-test and the Welch test discussed in section 2.1.3.

	policy 1	policy 2
task 1	normal distr. $m = 0$ $v = 1$	normal distr. $m = 0$ $v = 1$
task 2	normal distr. $m = 0$ $v = 1$	normal distr. $m = 0$ $v = 1$

TABLE 3.1: Normal data

The resulting actual significance level (estimated type I error probability) is shown in figure 14.

Within this simple scenario we note how most test perform extremely well even in the lower sample regime. The actual significance level is close

to the imposed nominal value of 0.05. This is expected as the normal distributed task comply with the assumptions of all the tests.

On the other hand, the stratified bootstrap test starts at an exceptionally high value of actual significance level (estimated type I error probability of roughly 20%) for lower sample size and quickly converges to a statistically significantly more conservative α. The convergence to a lower α should be expected and is the reason for having introduced inferential confidence intervals; descriptive confidence intervals are often too conservative when making inference [67]. Note the fact that when only a few samples are used, the stratified confidence intervals inference rejects the null hypothesis most of times; this result was previously reported on standard bootstrap confidence intervals by Colas et al. [18]. Evidently, with very limited data, CIs should not be used for inference. This is evident for both the stratified bootstrap IQM and mean results, although the mean appears to be somewhat less conservative, especially at lower sample size. Indeed IQM is calculated based on the two central quartiles of the sample which effectively means we consider only 50% of the data so we can expect it to be more conservative at lower sample size. The predictions through stratified bootstrap CI stabilize from a sample size of roughly 40.

In this first simulation, the two samples are drawn from two identical distributions which eliminates the difference between two tasks: in other words this scenario is exactly the same as the scenario with only one task.

We proceed with introducing differences between tasks. This can be obtained in two ways: by changing the distribution parameters or by changing the distribution family. Within this section we focus on the former approach, whereas the latter is discussed in section 3.4.
Let us firstly introduce a difference between the tasks by enforcing different means for different tasks. This is represented in table 3.2.

	policy 1	policy 2
task 1	normal distr. $m = 1$ $v = 1$	normal distr. $m = 1$ $v = 1$
task 2	normal distr. $m = 10$ $v = 1$	normal distr. $m = 10$ $v = 1$

TABLE 3.2: Normally distributed data with different mean

The following shall be noted:

- The two policies distribution are identical and we therefore do not expect any statistical significant difference between the two. Similarly to what discussed previously, tasks share the same distribution across policies.
- Contrarily to the data of table 3.1, the two tasks have a large difference in mean value: this is representative of the real scenarios encountered in

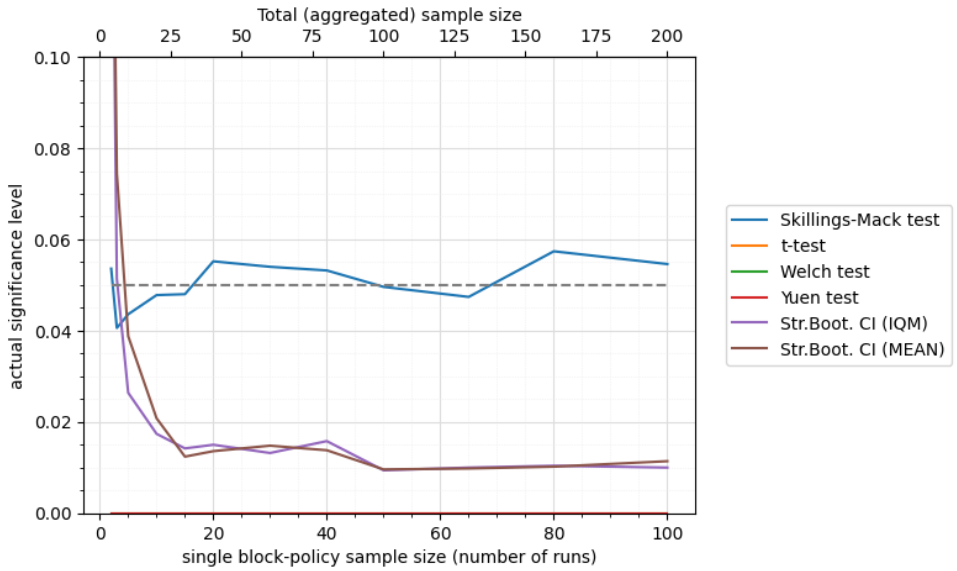


FIGURE 15: Actual significance level vs sample size of simulation with 5000 replications with two normally distributed tasks with different mean and same variance. Nominal $\alpha = 0.05$.

RL literature as tasks end-scores, even when normalized against human scores, are fundamentally different from task to task.

- The results, shown in figure 15, show that the Welch, Yuen and t tests never reject the null hypothesis so that their actual significance levels are overlapping with the bottom axis of the plot at $\alpha = 0$.

Note that when end-scores are aggregated from different tasks, as is commonly done in literature with the t-test, the Welch test or the bootstrap confidence intervals, the resulting test often fails to detect difference between policies. This is clearly evident in figure 15 where standard two sample test methods based on aggregated data fail to match with the nominal α and result in an actual significance level equal to 0. This can be interpreted as follows: given the increased variability of the aggregated sample, these methods do not recognize any difference between the two policies and are therefore too conservative.
Stratified bootstrap CI on the other hand show a trend similar to the one reported in figure 15 and appears not to be impacted by the different task mean.

3.4 Homogeneous tasks

This section introduces another source of variability: non-normal tasks.

Non-normal tasks We start by keeping the same number of tasks (i.e. two tasks) and introducing the *lognorm* and the *t* distributions as highlighted in

table 3.3.

	policy 1	policy 2
task 1	lognorm distr. $m = 0$ $v = 1$	lognorm distr. $m = 0$ $v = 1$
task 2	t distr. $m = 0$ $v = 1$	t distr. $m = 0$ $v = 1$

TABLE 3.3: Non-normally distributed data (lognorm and t distribution) with same mean and variance

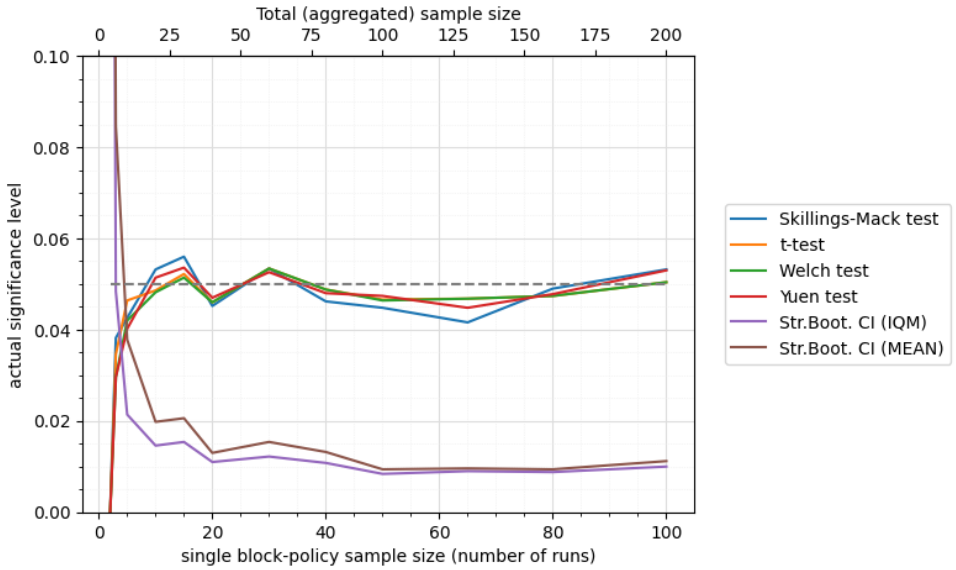


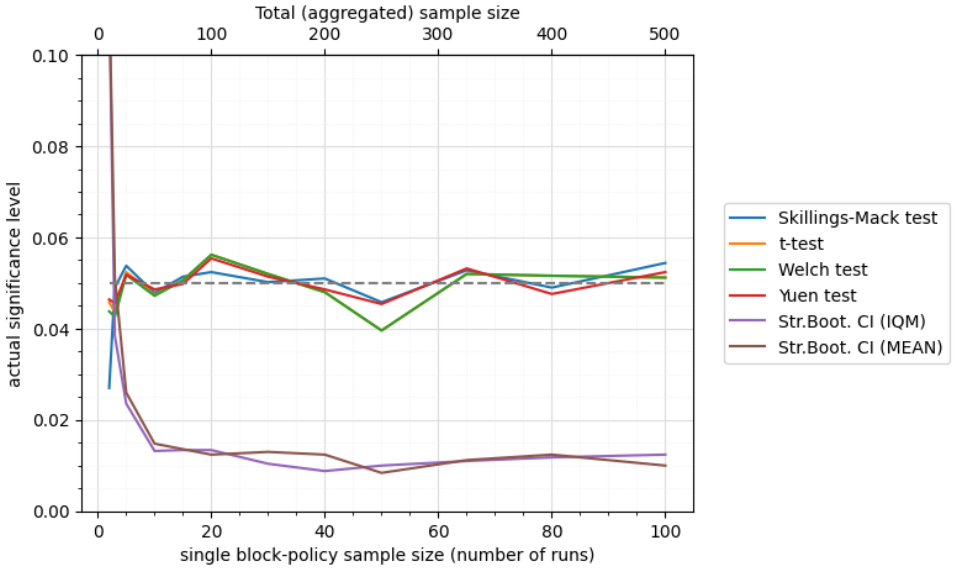
FIGURE 16: Actual significance level vs sample size of simulation with 5000 replications with two non-normally distributed tasks with same mean and variance. Nominal $\alpha = 0.05$.

The resulting graph is shown in figure 16. With respect to figure 14 we note how standard two sample test methods are more conservative with small sample size. This holds also for the the Skillings-Mack test which has no evident advantage in this scenario. The rather low actual significance level visible at low sample size is probably due to the tasks being not normally distributed.
At higher sample size, the actual significance level does not vary significantly although the use of non-normal distributions contradicts the hypotheses of most statistical tests. This shows a certain robustness of these methods against aggregating different task distributions.

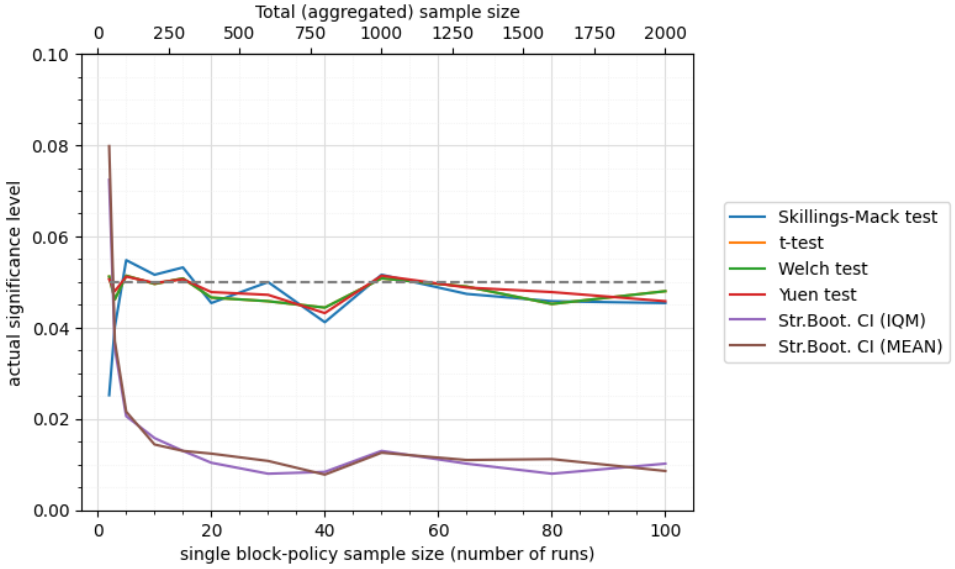
Increased number of tasks We now allow the number of tasks to increase from 2 to 5; we keep the complexities introduces in the previous paragraph as highlighted in table 3.4. Note the use of the normal, lognorm, t and the beta distribution. The resulting plot is shown in figure 17a. Interestingly the initial actual significance level calculated for the stratified bootstrap method

	policy 1	policy 2
task 1	norm distr. $m = 0$ $v = 1$	norm distr. $m = 0$ $v = 1$
task 2	norm distr. $m = 0$ $v = 1$	norm distr. $m = 0$ $v = 1$
task 3	lognorm distr. $m = 0$ $v = 1$	lognorm distr. $m = 0$ $v = 1$
task 4	t distr. $m = 0$ $v = 1$	t distr. $m = 0$ $v = 1$
task 5	beta distr. $m = 0$ $v = 1$	beta distr. $m = 0$ $v = 1$

TABLE 3.4: Five tasks homogeneously and non-normally distributed with same mean and variance



(A) Five non-normally distributed tasks with same mean and variance.



(B) Twenty non-normally distributed tasks with same mean and variance.

FIGURE 17: Actual significance level vs sample size of simulation with 5000 replications with 5 and 20 tasks. Nominal $\alpha = 0.05$.

is significantly lower than previously reported. The reason for this may be found in the increased sample size which is being used by the method: indeed stratified bootstrap takes one sample from each strata (i.e. each task) so having increased the number of tasks to five significantly impacted the simulation's minimum sample size. All methods benefit from the augmented sample size introduces by adding more tasks and show better performance in the low-sample-size regime.

Remark. The results for number of runs equal to 2 (figure 17a) should be compared with results of number of runs equal to 5 from figure 16.

Similarly to what reported in the previous paragraph, the methods provide reliable results even when their hypotheses do not hold. Increasing even further the number of tasks further has no significant effect for higher number of runs. This is shown in figure 17b which duplicates the data of table 3.4to obtain 20 tasks total, all the rest being equal. Notably, as the number of tasks increases (so the total aggregated sample size), stratified bootstrap CIs perform better at lower number of runs (i.e. single block-policy sample size).

3.5 Heterogeneous tasks

We finally introduce perturbations within policies. This is performed in two ways: by introducing a different variance between tasks or by allowing task distribution to vary across policies.

Task variance We start by introducing a different variance across policies. The assumed data is non-normally distributed and with variances as shown in table 3.4. Note that variance changes both across tasks and across policies.

	policy 1	policy 2
task 1	norm distr. $m = 0$ $v = 1$	norm distr. $m = 0$ $v = 1$
task 2	norm distr. $m = 0$ $v = 2$	norm distr. $m = 0$ $v = 4$
task 3	lognorm distr. $m = 0$ $v = 10$	lognorm distr. $m = 0$ $v = 12$
task 4	t distr. $m = 0$ $v = 0.1$	t distr. $m = 0$ $v = 0.5$
task 5	beta distr. $m = 0$ $v = 8$	beta distr. $m = 0$ $v = 6$

TABLE 3.5: Five tasks homogeneously and non-normally distributed with same mean. Variance changes across tasks and across policies.

The resulting plot is shown in figure 18. The results are aligned with the ones presented previously; the different test methods seem robust against changes in the variance both across policies and across tasks.

Distributions This paragraph introduces different task distribution across policies. Table 3.6 presents the distributions used and their mean and variance. In bold we highlight differences between policies. The introduction of

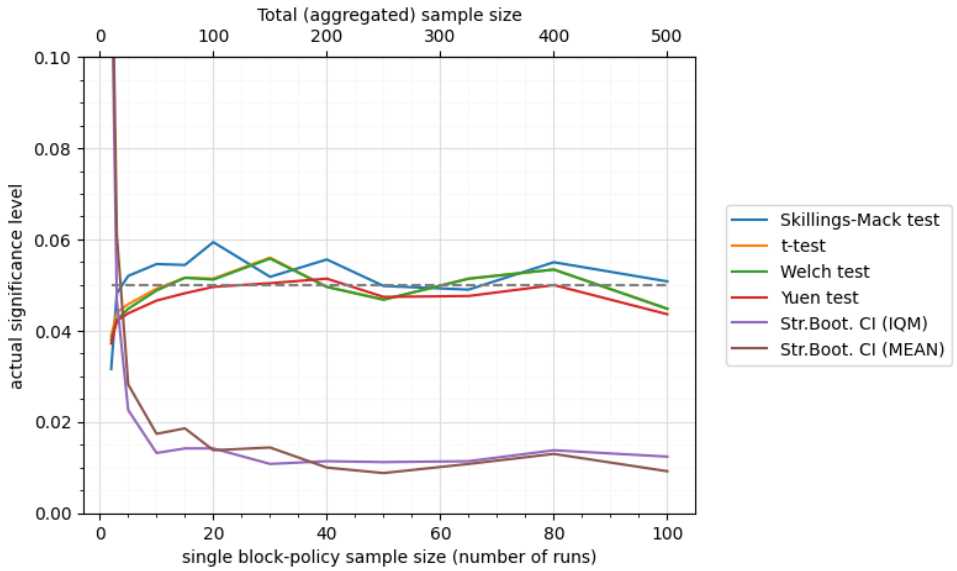


FIGURE 18: Actual significance level vs sample size of simulation with 5000 replications with five tasks homogeneously and non-normally distributed with same mean. Variance changes across tasks and across policies. Nominal $\alpha = 0.05$.

differences between policies is a recognized criticality in RL as discussed in section 2.1.4.

	policy 1	policy 2
task 1	**norm** distr. $m = 0$ $v = 1$	**gamma** distr. $m = 0$ $v = 1$
task 2	**norm** distr. $m = 0$ $v = 1$	**lognorm** distr. $m = 0$ $v = 1$
task 3	**lognorm** distr. $m = 0$ $v = 1$	**dweibul** distr. $m = 0$ $v = 1$
task 4	t distr. $m = 0$ $v = 1$	t distr. $m = 0$ $v = 1$
task 5	beta distr. $m = 0$ $v = 1$	beta distr. $m = 0$ $v = 1$

TABLE 3.6: Five tasks heterogeneously and non-normally distributed with same mean and variance. Highlighted differences between policies

Results are provided within figure 19 which shows a positive onward trend for the Skillings-Mack test, the Yuen test and the stratified bootstrap IQM. This is especially evident at higher sample size whereas for lower sample size the diffences are negligible. This may be explained with the fact that this scenario violates the assumption 2 of section 2.3.1, i.e. that the different policies have the same distribution except for the additive effect of the policy and of the block or task.

3.6 Statistical Power

This section focuses on statistical power. As mentioned in section 3.1 statistical power is related to the type-II error probability (i.e. the probability of

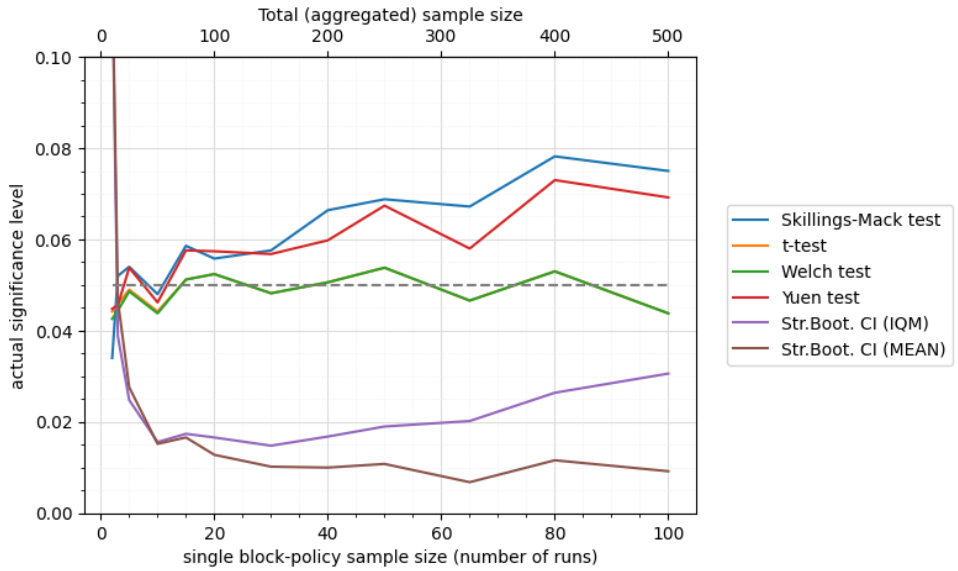


FIGURE 19: Actual significance level vs sample size of simulation with 5000 replications with five tasks heterogeneously and non-normally distributed with same mean and variance. Nominal $\alpha = 0.05$.

incorrectly accepting the null hypothesis when the alternative holds) by the formula $1 - \beta$ with β being the type-II error probability.
To this aim we enforce the alternative hypothesis by allowing the two policies to have a controlled different mean, Δ: this value is critical, as larger differences lead to an easier inference. In order to evaluate the impact of this parameter we will assume the following three mean increases: $\Delta = 0.5$, $\Delta = 1$ and $\Delta = 2$. The parameter Δ is added to each task mean during comparison as shown in table 3.7: in this way, when $\Delta \neq 0$, the alternative hypothesis is enforced. The use of different values for Δ allows us to evaluate the effect of small to large effects.

	policy 1	policy 2
task 1	normal distr. $m = 0$ $v = 1$	normal distr. $m = 0 + \Delta$ $v = 1$
task 2	normal distr. $m = 0$ $v = 1$	normal distr. $m = 0 + \Delta$ $v = 1$

TABLE 3.7: Statistical power: task distributions share the same mean.

Figure 20a shows the resulting three plots for varying Δ. We note the following:

- As expected with increasing values of Δ, the statistical power of all tests increases. Similarly the power increases with increasing sample size for fixed Δ.
  For high values of Δ all methods align and show high power.

- Similarly to the findings of previous section, the Skillings-Mack test results are aligned with other 2 sample tests considered within the simulation.

- Stratified bootstrap confidence intervals are shown to be rather conservative most of times: this is especially visible in the left pane ($\Delta = 0.5$) where stratified bootstrap CI shows a 20% absolute drop in power w.r.t. other methods.
  For higher values of Δ, the performances of stratified bootstrap CI are aligned with the 2 sample tests and performs slightly better in the low sample regime with a starting power of roughly 80% against a power of roughly 60% for 2-sample tests.

In order to appreciate the improvement introduced by the Skillings-Mack test we allow the tasks mean to vary. We keep using the same parameter Δ for encoding the difference between policies as shown in table 3.8.

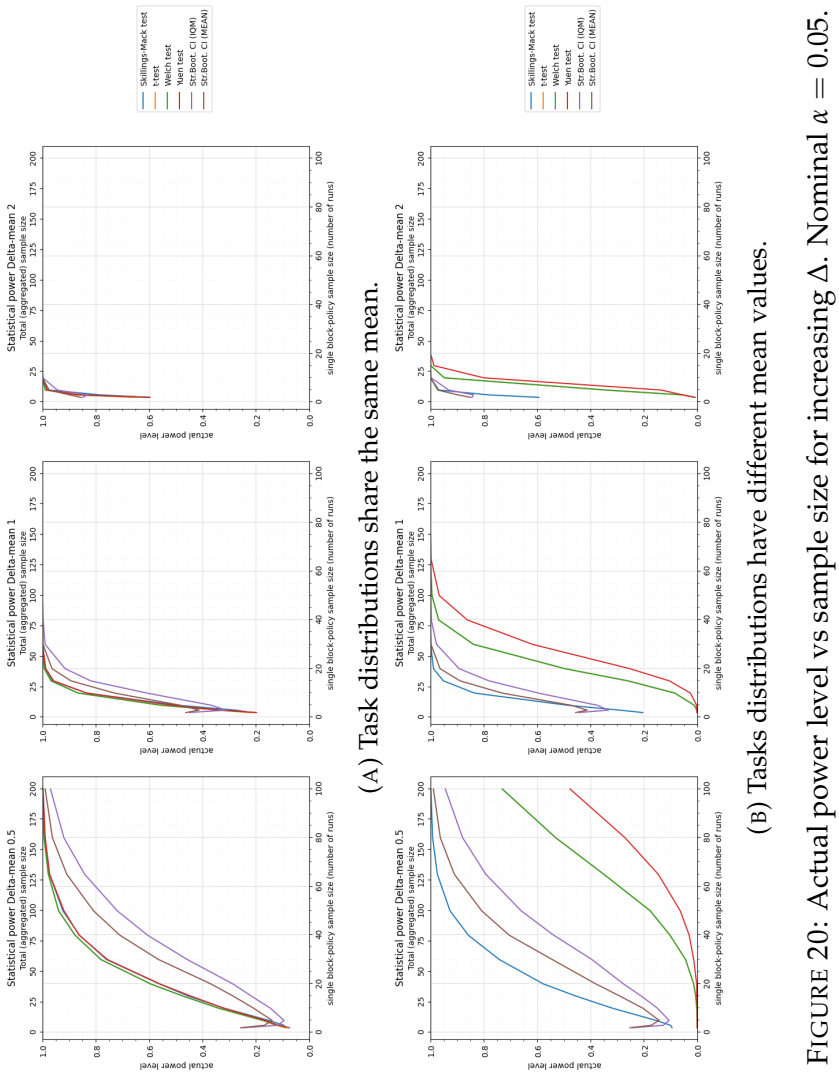
	policy 1	policy 2
task 1	normal distr. $m = 6$ $v = 1$	normal distr. $m = 6 + \Delta$ $v = 1$
task 2	normal distr. $m = 10$ $v = 1$	normal distr. $m = 10 + \Delta$ $v = 1$

TABLE 3.8: Statistical power: tasks distributions have different mean values.

With respect to the distributions shown earlier we have decreased the difference between the mean of the two tasks from 1-10 to 6-10: this change allows to better appreciate the transition between $\Delta = 0.5$ to $\Delta = 2$.
The resulting plot is shown in figure 20b where we note that:

- The Skillings-Mack test and the stratified bootstrap confidence interval method have kept a similar performance to what shown in figure 3.7;

- Similarly to what was observed in figure 15 the Yuen, Welch and the t tests are significavely affected by the difference in mean between tasks with statistical power dropping. Evidently, when a strong effect is present ($\Delta = 2$) also these methods are able to detect the change but this is not true for smaller effects ($\Delta = 0.5$).



(A) Task distributions share the same mean.

(B) Tasks distributions have different mean values.

FIGURE 20: Actual power level vs sample size for increasing Δ. Nominal $\alpha = 0.05$.

3.7 Conclusions

This chapter has showed the application of the Skillings-Mack test on synthetic data and its comparison against other methods.
From sections 3.3, 3.4 and 3.5 it is evident that the test performs similarly to most other 2 sample tests under most situations with one notable exceptions: when different tasks distribution have a different mean value, the skillings-mack test is the only two sample test to provide correct results. This result is confirmed by the statistical power calculations of section 3.6 where the skillings-mack test outperforms all other test methods, including the stratified bootstrap CIs estimates. These considerations naturally hold also for the inferential confidence intervals built on top of the skillings-mack test which aligns with the latter under all conditions.

Based on these results we recommend using the skillings-mack test for policy comparison. Thanks to this approach it is possible to reduce the number of points for each sample to roughly 20 which corresponds to 1 runs required for each task instead of the suggested 3-10 on the Atari 100k benchmark [2];

Appendix A

Appendix A - Atari 100k distributions

Hereafter we provide for each game of the Atari 100K the distribution plot and best fit resulting from the data [1].

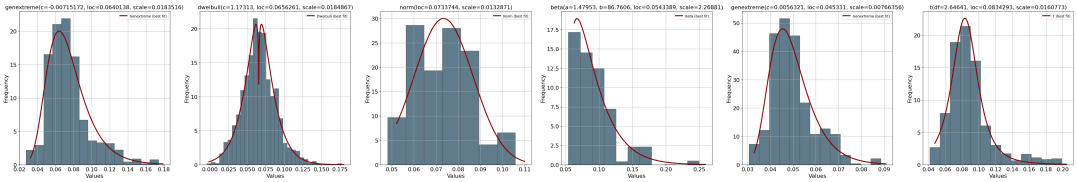


FIGURE A.1: Alien fit

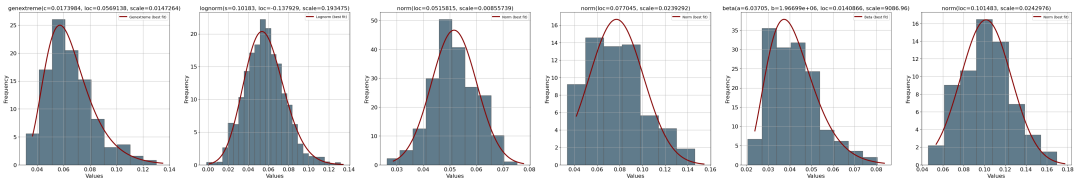


FIGURE A.2: Amidar fit

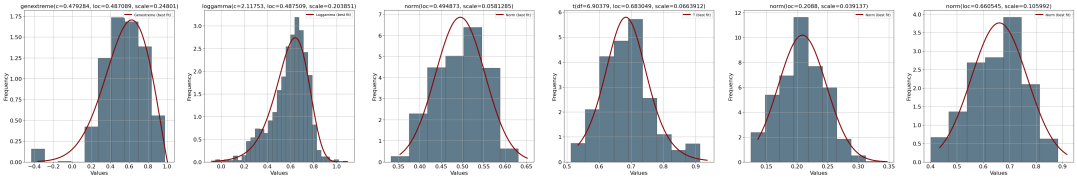


FIGURE A.3: Assault fit

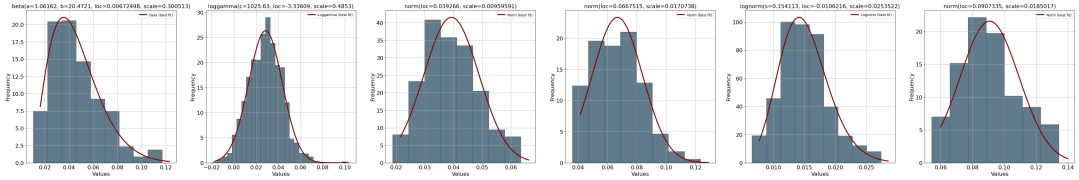


FIGURE A.4: Asterix fit

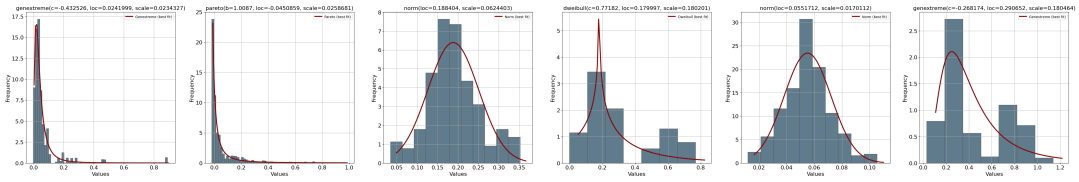


FIGURE A.5: BankHeist fit

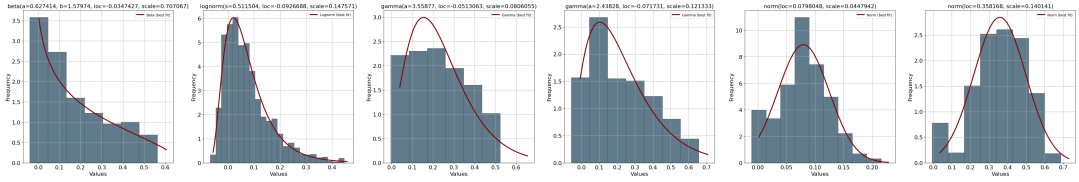


FIGURE A.6: BattleZone fit

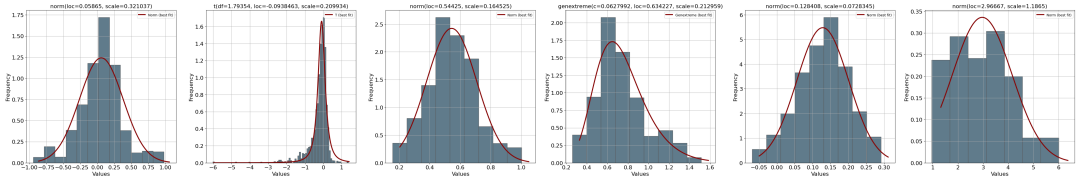


FIGURE A.7: Boxing fit

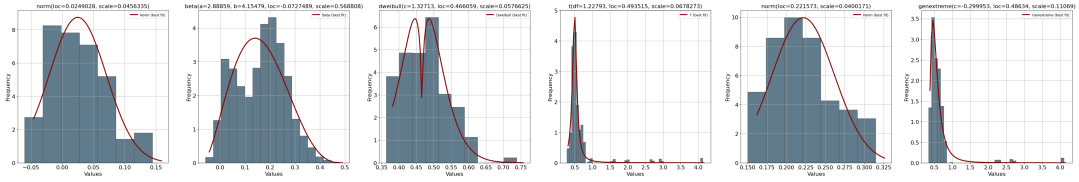


FIGURE A.8: Breakout fit

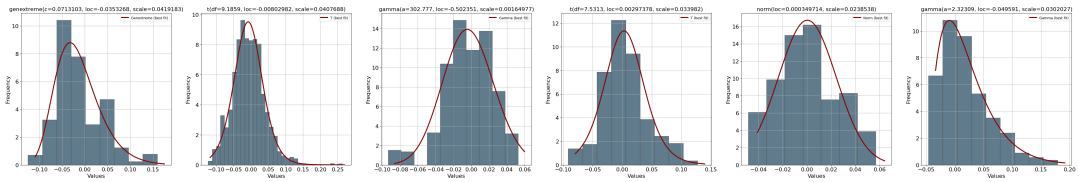


FIGURE A.9: Chopper Command fit

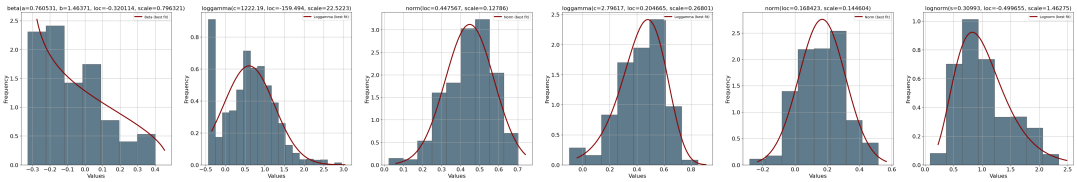


FIGURE A.10: Crazy Climber fit

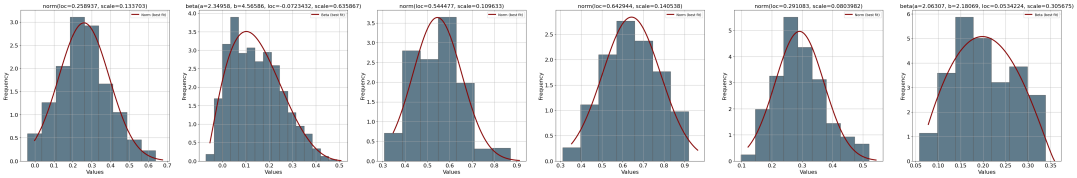


FIGURE A.11: Deamon Attack fit

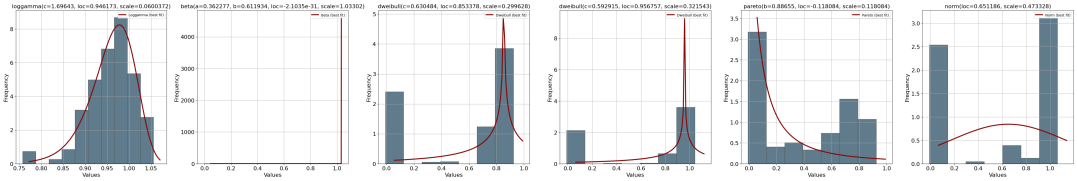


FIGURE A.12: Freeway fit

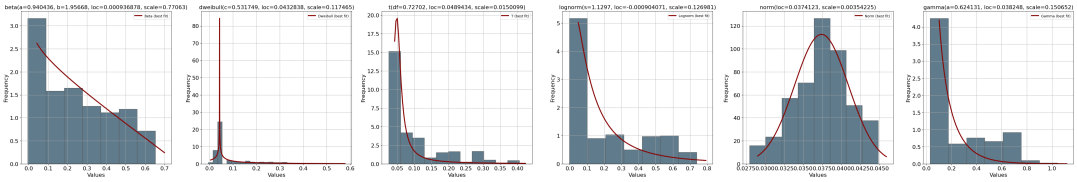


FIGURE A.13: Frostbite fit

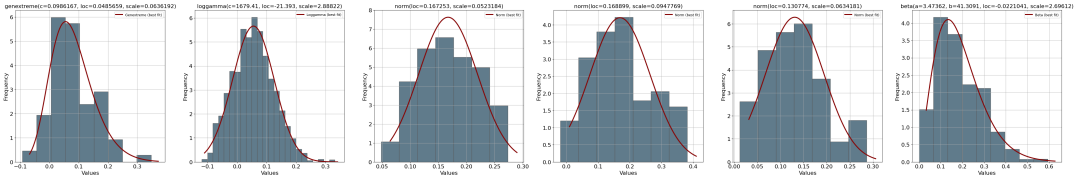


FIGURE A.14: Gopher fit

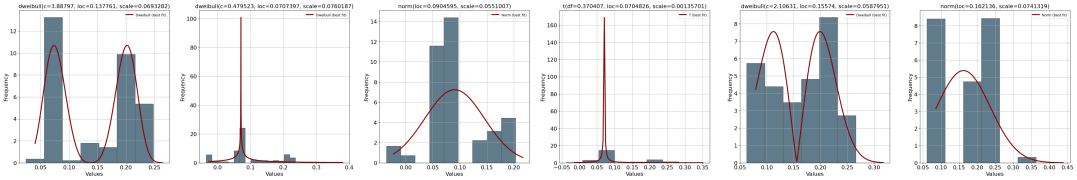


FIGURE A.15: Hero fit

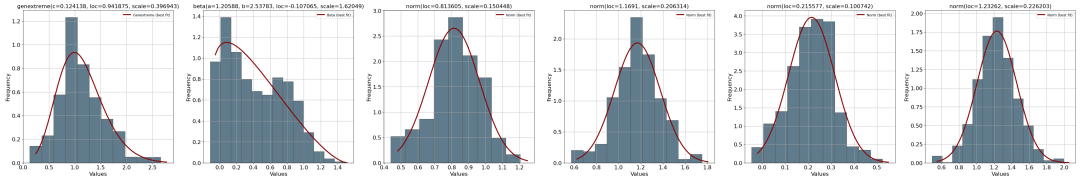


FIGURE A.16: Jamesbond fit

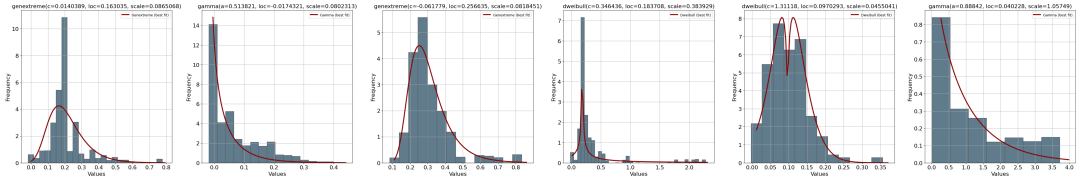


FIGURE A.17: Kangaroo fit

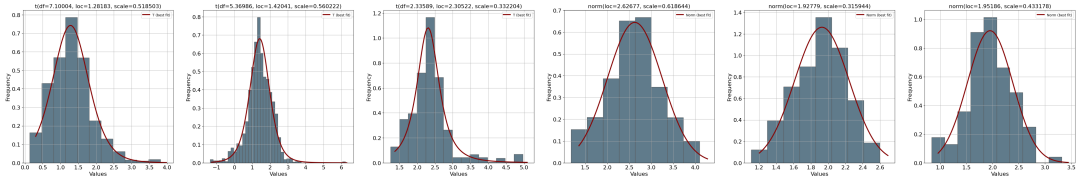


FIGURE A.18: Krull fit

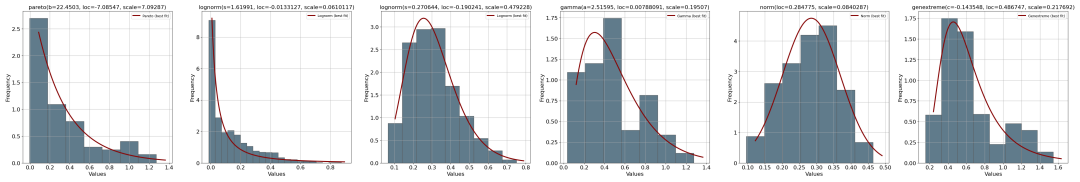


FIGURE A.19: Kung Fu Master fit

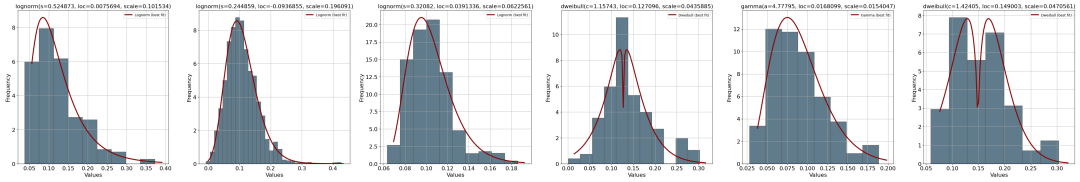


FIGURE A.20: MsPacman fit

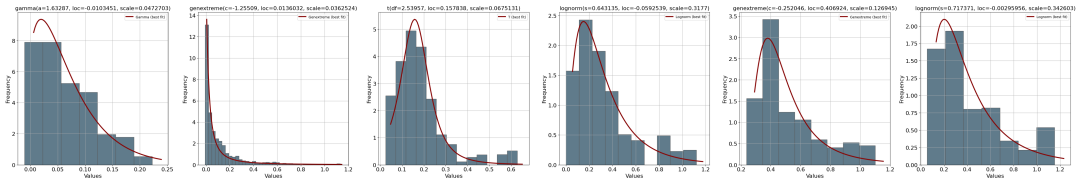


FIGURE A.21: Pong fit

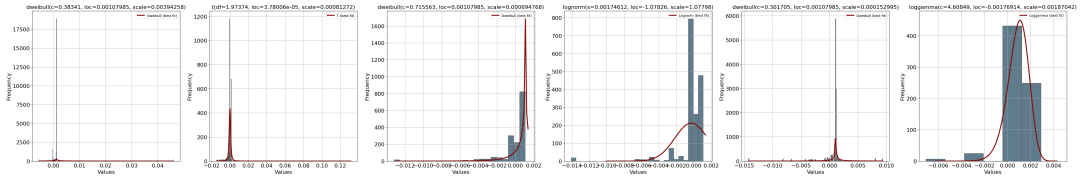


FIGURE A.22: Private Eye fit

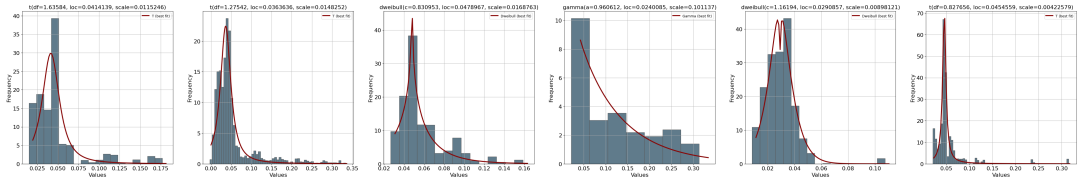


FIGURE A.23: Qbert fit

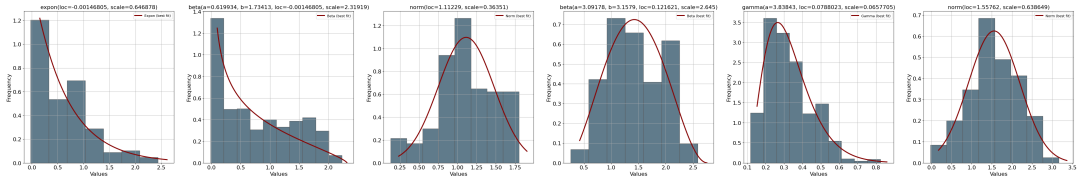


FIGURE A.24: Road Runner fit

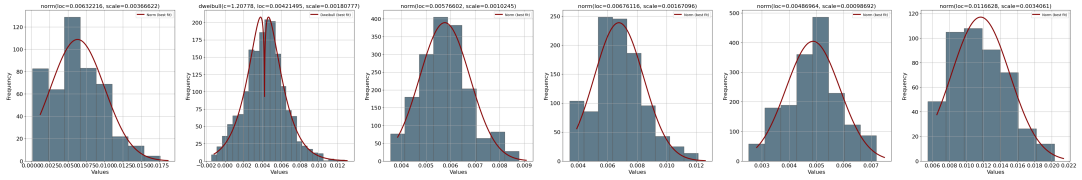


FIGURE A.25: Seaquest fit

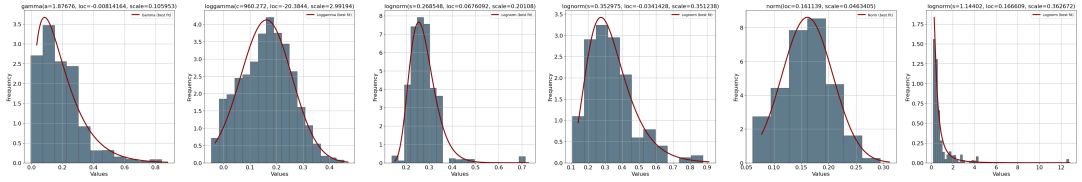


FIGURE A.26: UpNDown fit

Appendix B

Appendix B - Python code

This section contains the main Python code used through this work.

Section B.1 shows code to perform Skillings-Mack test critical value calculation under the assumption of equal number of replications $c_{ij} = c$ (refer to section 2.3.2. The code is heavily inspired by the cMackSkil function of the R NSM3 package [34] and has been tested on Python 3.10 and 3.11.

Section B.2 shows an extract of the code used for running actual significance level and statistical power simulations.

B.1 Skillings-Mack test critical value computation

```
def cMackSkil(alpha, k, n, c, method=None, n_mc=10000):
    # Compute the Skillings Mack test statistic critical value.
    # Holds for the case of complete block
    # with equal number of replications c.
    outp = {}
    outp["stat.name"] = "Mack-Skillings MS"
    outp["n.mc"] = n_mc
    outp["k"] = k  # Number of policies or treatments
    outp["n"] = n  # Number of tasks or blocks
    outp["c"] = c  # Number of replications on each block

    if alpha > 1 or alpha < 0 or not isinstance(alpha, (int, float)):
        raise ValueError("Error: Check alpha value!")

    outp["alpha"] = alpha

    num_obs = outp["k"] * outp["n"] * outp["c"]
    # Total number of observations

    if method is None:
        if np.math.factorial(outp["c"] * outp["k"]
                             * outp["n"]) <= 100000:
            # Exact computation is feasible
            method = "Exact"
        else:
```

```
        # Exact computation is computationally prohibitive
        method = "Monte Carlo"
outp["method"] = method

def MS_calc(obs_data):
    S_vec = [
        np.sum(obs_data[:,
            (i - 1) * outp["c"] : (i * outp["c"])]) / outp["c"]
        for i in range(1, outp["k"] + 1)
    ]
    # MS test statistic
    MS_stat = 12 / (outp["k"] * (num_obs + outp["n"])) * np.sum(
        np.array(S_vec) ** 2
    ) - 3 * (num_obs + outp["n"])
    return MS_stat

if outp["method"] == "Exact":
    possible_ranks = np.tile(
        np.array(range(1, outp["c"] * outp["k"] + 1)),
                 [outp["n"], 1]
    )
    possible_perm = [
        np.reshape(arr, possible_ranks.shape)
        for arr in list(permutations(
            np.reshape(possible_ranks, -1)))
    ]  # All possible permutations of ranks
    exact_dist = list(map(MS_calc, list(possible_perm)))

    MS_vals = np.unique(exact_dist)
    MS_probs = np.array([np.sum(exact_dist == val)
        for val in MS_vals]) /
    ( np.math.factorial(outp["c"] * outp["k"] * outp["n"]) )
    MS_dist = np.column_stack((MS_vals, MS_probs))
    upper_tails = np.column_stack(
        (np.flip(MS_dist[:, 0]),
            np.cumsum(np.flip(MS_dist[:, 1])))
    )
    outp["cutoff_U"] = upper_tails[np.max(
        np.where(upper_tails[:, 1] <= alpha)), 0]
    outp["true_alpha_U"] = upper_tails[
        np.max(np.where(upper_tails[:, 1] <= alpha)), 1
    ]

if outp["method"] == "Monte Carlo":
    possible_ranks = np.tile(
        np.array(range(1, outp["c"] * outp["k"] + 1)),
                 [outp["n"], 1]
```

```
        )
        mc_perm = np.zeros((outp["n"], outp["c"] * outp["k"]))
        mc_stats = np.zeros(outp["n.mc"])
        for i in range(outp["n.mc"]):
            for j in range(outp["n"]):
                # Instead of calculating all possible permutations,
                # we take a random permutation of ranks
                # and repeat n.mc times
                mc_perm[j, :] = np.random.permutation(
                    possible_ranks[j, :])
            mc_stats[i] = round(MS_calc(mc_perm), 5)

        mc_vals = np.unique(mc_stats)
        mc_dist = np.array([np.sum(mc_stats == val)
            for val in mc_vals]) / outp["n.mc"]

        upper_tails = np.column_stack((np.flip(mc_vals),
            np.cumsum(np.flip(mc_dist))))
        outp["cutoff_U"] = upper_tails[np.max(np.where(
            upper_tails[:, 1] <= alpha)), 0]
        outp["true_alpha_U"] = upper_tails[
            np.max(np.where(upper_tails[:, 1] <= alpha)), 1
        ]

    if outp["method"] == "Asymptotic":
        outp["p_val"] = chi2.ppf(1 - alpha, outp["k"] - 1)

    return outp
```

B.2 Significance level and Statistical Power simulation

```
Cs = list([2, 3, 5, 10, 15, 20, 30, 40, 50, 65, 80, 100])


def run_simulation(
    nominal_alpha: float = 0.05,
    K: int = 2,
    N: int = 5,
    Nsim: int = 100,
    TASK_MEAN: list = list([[0, 0, 0, 0, 0], [0, 0, 0, 0, 0]]),
    TASK_VAR: list = list([[1, 1, 1, 1, 1], [1, 1, 1, 1, 1]]),
    TASK_DISTR: list = list(
        [
            ["norm", "norm", "norm", "norm", "norm"],
            ["norm", "norm", "norm", "norm", "norm"],
```

```
        ]
    ),
    SB_reps: int = 100,
):
    # K Number of treatments (i.e. number of algorithms)
    # N Number of blocks (i.e. number of games)
    # Nsim Number of simulation points for computing
    # the effective alpha
    # task_mean list of means for each task.
    # The default value assumes
    # the two policies have the same distribution.
    # task_var list of variances for each task.
    # The default value assumes
    # the two policies have the same distribution.
    # task_distr list of distribution names for each task.
    # The default value assumes the two policies
    # have the same distribution.

    if np.shape(np.array(TASK_MEAN)) != (K, N):
        raise ValueError("Wrong shape length")

    if np.shape(np.array(TASK_MEAN)) !=
        np.shape(np.array(TASK_VAR))
        and np.shape( np.array(TASK_MEAN)
    ) != np.shape(np.array(TASK_DISTR)):
        raise ValueError(
            "The shape of TASK_MEAN,
                TASK_VAR and TASK_DISTR do not correspond"
        )

    MSv = np.zeros(Nsim)
    p_vals_t = np.zeros(Nsim)
    p_vals_welch = np.zeros(Nsim)
    p_vals_welch_trim = np.zeros(Nsim)
    SBv = np.zeros(Nsim)
    SBov = np.zeros(Nsim)
    SBv_m = np.zeros(Nsim)
    SBov_m = np.zeros(Nsim)

    Cs = list([2, 3, 5, 10, 15, 20, 30,
               40, 50, 65, 80, 100])

    fpos_MS = np.zeros(len(Cs))
    fpos_t = np.zeros(len(Cs))
    fpos_welch = np.zeros(len(Cs))
    fpos_welch_trim = np.zeros(len(Cs))
    fpos_SB = np.zeros(len(Cs))
```

```
fpos_SBo = np.zeros(len(Cs))
fpos_SB_m = np.zeros(len(Cs))
fpos_SBo_m = np.zeros(len(Cs))

for g in range(len(Cs)):
    # number of runs, i.e. single block-policy sample size
    C = Cs[g]
    num_obs = K * N * C
    critical_value = cMackSkil(nominal_alpha, K, N, C,
                               method="Asymptotic")["p_val"]
    for f in range(Nsim):  # Repeat the experiment Nsim times.
        observ = np.zeros((N, K * C))
        for i in range(N):  # Loop over the number of tasks
            for j in range(K):  # Loop over the different policies
                task_mean = TASK_MEAN[j][i]
                task_var = TASK_VAR[j][i]
                task_distr = TASK_DISTR[j][i]
                match j:
                    case 0:  # Reference policy/treatment
                        ref_sample = rvs_enforce(
                            distr_name=task_distr,
                            mean=task_mean,
                            var=task_var,
                            size=C,
                        )
                        observ[i, j * C : j * C + C] = ref_sample
                    case 1:  # Alternative policy/treatment
                        altern_sample = rvs_enforce(
                            distr_name=task_distr,
                            mean=task_mean,
                            var=task_var,
                            size=C,
                        )
                        observ[i, j * C : j * C + C] = altern_sample

        # Compute the Mack-Skillings test
        ranks = stats.rankdata(
            observ, axis=1
        )  # rank 1 corresponds to the smallest value
        ranks = -np.add(ranks, -np.max(ranks + 1))
        # rank 1 is the best algorithm
        avg_rank = np.zeros((N, K))
        for i in range(N):
            for j in range(K):
                avg_rank[i, j] = np.mean(ranks[i, j * C : j * C + C])
        S = np.mean(avg_rank, axis=0) * N
        MS = 12 / (K * (num_obs + N)) * np.sum(np.power(S, 2)) -
```

```
    3 * ( num_obs + N  )  # MS statistic score
MSv[f] = MS

# Compute 2-sample tests (t, welch yuen).
#Assuming we merge all the data from
# the different tasks together
sample1 = np.reshape(observ[:, 0:C], -1)
sample2 = np.reshape(observ[:, C : 2 * C], -1)
_, p_vals_t[f] = stats.ttest_ind(
    sample1, sample2, equal_var=True,
    alternative="two-sided", trim=0
)  # t test
_, p_vals_welch[f] = stats.ttest_ind(
    sample1, sample2, equal_var=False,
    alternative="two-sided", trim=0
)  # Welch test
_, p_vals_welch_trim[f] = stats.ttest_ind(
    sample1, sample2, equal_var=False,
    alternative="two-sided", trim=0.1
)  # Yuen test

# Stratified bootstrap
# aggregate_interval_estimates contains the upper
# and lower bound of the CI
_, aggregate_interval_estimates =
    strat_bootstrap(observ, C, reps=SB_reps)
SBv[f] = intervals_overlap(
    aggregate_interval_estimates, 0
)  # stratified bootstrap CI overlap IQM
SBv_m[f] = intervals_overlap(
    aggregate_interval_estimates, 1
)  # stratified bootstrap CI overlap MEAN

# Stratified Boostrap ICIs.
MSchi2 = 1 - stats.chi2.cdf(MS, K - 1)
# p-value of the MS statistic

# by default we keep the ICIs equal to the CI.
eps = 1

# IQM implementation
m1 = IQM(np.transpose(observ[:, 0:C]))
m2 = IQM(np.transpose(observ[:, C : C + C]))
if MS >= critical_value:
    if m2 >= m1:
        eps = (
            (m2 - m1)
```

```
                * (1 + MSchi2 - nominal_alpha)
                / (
                    m2
                    - m1
                    + (
                        aggregate_interval_estimates[0][1, 0]
                        - aggregate_interval_estimates[1][0, 0]
                    )
                )
            )  # Uses MackSkilling. Assumes IQM is the first
        else:
            eps = (
                (m1 - m2)
                * (1 + MSchi2 - nominal_alpha)
                / (
                    m1
                    - m2
                    + aggregate_interval_estimates[1][1, 0]
                    - aggregate_interval_estimates[0][0, 0]
                )
            )  # Uses MackSkilling. Assumes IQM is the first
        eps = 0.999 * np.min(
            [eps, 1]
        )  # 0.999 is there to account for any numerical error.
    elif (
        intervals_overlap(aggregate_interval_estimates, 0) < 0
    ):  # the null is not rejected at alpha,
    # this means that the intervals must overlap.
    # If not, we have to fix it.
        if m2 >= m1:
            eps = (
                (m2 - m1)
                * (1 + MSchi2 - nominal_alpha)
                / (
                    m2
                    - m1
                    + (
                        aggregate_interval_estimates[0][1, 0]
                        - aggregate_interval_estimates[1][0, 0]
                    )
                )
            )  # Uses MackSkilling. Assumes IQM is the first
        else:
            eps = (
                (m1 - m2)
                * (1 + MSchi2 - nominal_alpha)
                / (
```

```
                    m1
                    - m2
                    + aggregate_interval_estimates[1][1, 0]
                    - aggregate_interval_estimates[0][0, 0]
                )
            )  # Uses MackSkilling. Assumes IQM is the first
        eps = eps * 1.001
        # 1.001 is there to account for any numerical error.

SBov[f] = intervals_overlap_inference(
    aggregate_interval_estimates, m1, m2, eps, 0
)

# MEAN implementation
eps = 1  # by default we keep the ICIs equal to the CI.
# Calculate the center of the CI.
m1 = MEAN(
    np.transpose(observ[:, 0:C])
)  # Maps the implementation of the original paper
m2 = MEAN(
    np.transpose(observ[:, C : C + C])
)  # Maps the implementation of the original paper
if MS >= critical_value:
    if m2 >= m1:
        eps = (
            (m2 - m1)
            * (1 + MSchi2 - nominal_alpha)
            / (
                m2
                - m1
                + (
                    aggregate_interval_estimates[0][1, 1]
                    - aggregate_interval_estimates[1][0, 1]
                )
            )
        )  # Uses MackSkilling. Assumes IQM is the first
    else:
        eps = (
            (m1 - m2)
            * (1 + MSchi2 - nominal_alpha)
            / (
                m1
                - m2
                + aggregate_interval_estimates[1][1, 1]
                - aggregate_interval_estimates[0][0, 1]
            )
        )  # Uses MackSkilling. Assumes IQM is the first
```

```
            eps = 0.999 * np.min(
                [eps, 1]
            )  # 0.999 is there to account for any numerical error.
        elif (
            intervals_overlap(aggregate_interval_estimates, 1) < 0
        ):  # the null is not rejected at alpha,
        # this means that the intervals must overlap.
        # If not, we have to fix it.
            if m2 >= m1:
                eps = (
                    (m2 - m1)
                    * (1 + MSchi2 - nominal_alpha)
                    / (
                        m2
                        - m1
                        + (
                            aggregate_interval_estimates[0][1, 1]
                            - aggregate_interval_estimates[1][0, 1]
                        )
                    )
                )  # Uses MackSkilling. Assumes IQM is the first
            else:
                eps = (
                    (m1 - m2)
                    * (1 + MSchi2 - nominal_alpha)
                    / (
                        m1
                        - m2
                        + aggregate_interval_estimates[1][1, 1]
                        - aggregate_interval_estimates[0][0, 1]
                    )
                )  # Uses MackSkilling. Assumes IQM is the first
            eps = eps * 1.001
            # 1.001 is there to account for any numerical error.

        SBov_m[f] = intervals_overlap_inference(
            aggregate_interval_estimates, m1, m2, eps, 1
        )

    fpos_MS[g] = np.sum(MSv >= critical_value) / Nsim
    fpos_t[g] = np.sum(p_vals_t <= nominal_alpha) / Nsim
    fpos_welch[g] = np.sum(p_vals_welch
                           <= nominal_alpha) / Nsim
    fpos_welch_trim[g] = np.sum(p_vals_welch_trim
                                <= nominal_alpha) / Nsim
    fpos_SB[g] = np.sum(SBv <= 0) / Nsim
    fpos_SBo[g] = np.sum(SBov <= 0) / Nsim
```

```
        fpos_SB_m[g] = np.sum(SBv_m <= 0) / Nsim
        fpos_SBo_m[g] = np.sum(SBov_m <= 0) / Nsim

    return (
        fpos_MS,
        fpos_t,
        fpos_welch,
        fpos_welch_trim,
        fpos_SB,
        fpos_SBo,
        fpos_SB_m,
        fpos_SBo_m,
    )
```

Bibliography

[1] Rishabh Agarwal et al. *Data for individual runs on Atari 100k, ALE, DM Control and Procgen*. https://console.cloud.google.com/storage/browser/rl-benchmark-data. [Online; accessed 01-June-2023].

[2] Rishabh Agarwal et al. "Deep reinforcement learning at the edge of the statistical precipice". In: *Advances in neural information processing systems* 34 (2021), pp. 29304–29320.

[3] Rishabh Agarwal et al. *Stratified Bootstrap Confidence Interval implementation*. https://bit.ly/statistical_precipice_colab. [Online; accessed 03-Apr-2024].

[4] Clarivate Analytics. *Web of Science Core Collection: Definition of Source Items*. https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-Definition-of-Source-Items. [Online; accessed 24-Mar-2024].

[5] Clarivate Analytics. *Web of Science core collection subject categories*. https://webofscience.help.clarivate.com/en-us/Content/wos-core-collection/wos-core-collection.html. [Online; accessed 14-Mar-2024].

[6] Clarivate Analytics. *Web of Science Search Rules - Lemmization and Stemming*. https://webofscience.help.clarivate.com/en-us/Content/search-rules.htm. [Online; accessed 14-Apr-2024].

[7] Massimo Aria and Corrado Cuccurullo. "bibliometrix: An R-tool for comprehensive science mapping analysis". In: *Journal of informetrics* 11.4 (2017), pp. 959–975.

[8] Massimo Aria and Corrado Cuccurullo. *Bibliometrix Resources*. https://bibliometrix.org/biblioshiny/assets/player/KeynoteDHTMLPlayer.html. [Online; accessed 24-Mar-2024].

[9] Kai Arulkumaran et al. "A brief survey of deep reinforcement learning". In: *arXiv preprint arXiv:1708.05866* (2017).

[10] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. "Neuronlike adaptive elements that can solve difficult learning control problems". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13.5 (1983), pp. 834–846. DOI: 10.1109/TSMC.1983.6313077.

[11] Marc G. Bellemare, Will Dabney, and Rémi Munos. "A Distributional Perspective on Reinforcement Learning". In: *CoRR* abs/1707.06887 (2017). arXiv: 1707.06887. URL: http://arxiv.org/abs/1707.06887.

[12] Marc G Bellemare et al. "The arcade learning environment: An evaluation platform for general agents". In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 253–279.

[13] Carlo Bonferroni. "Teoria statistica delle classi e calcolo delle probabilita". In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8 (1936), pp. 3–62.

[14] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series. Brooks/Cole Publishing Company, 1990. ISBN: 9780534119584. URL: https://books.google.it/books?id=nA_vAAAAMAAJ.

[15] Stephanie CY Chan et al. "Measuring the reliability of reinforcement learning algorithms". In: *arXiv preprint arXiv:1912.05663* (2019).

[16] Kaleigh Clary et al. "Let's Play Again: Variability of Deep Reinforcement Learning Agents in Atari Environments". In: *arXiv preprint arXiv:1904.06312* (2019).

[17] Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. "A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms". In: *arXiv preprint arXiv:1904.06979* (2019).

[18] Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. "How many random seeds? statistical power analysis in deep reinforcement learning experiments". In: *arXiv preprint arXiv:1806.08295* (2018).

[19] The SciPy community. *Scipy v.1.12.0 Manual*. https://docs.scipy.org/doc/scipy/reference/stats.html. [Online; accessed 28-Feb-2024].

[20] Ralph D'Agostino and Egon S Pearson. "Tests for departure from normality. Empirical results for the distributions of b 2 and sqrt(b)". In: *Biometrika* 60.3 (1973), pp. 613–622.

[21] Michael John De Smith. *Statistical analysis handbook*. The Winchelsea Press, 2018.

[22] Janez Demšar. "Statistical comparisons of classifiers over multiple data sets". In: *The Journal of Machine learning research* 7 (2006), pp. 1–30.

[23] Naveen Donthu et al. "How to conduct a bibliometric analysis: An overview and guidelines". In: *Journal of Business Research* 133 (2021), pp. 285–296. ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2021.04.070. URL: https://www.sciencedirect.com/science/article/pii/S0148296321003155.

[24] Bradley Efron. "Bootstrap methods: another look at the jackknife". In: *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 569–593.

[25] R. A. Fisher. "On the Mathematical Foundations of Theoretical Statistics". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222 (1922), pp. 309–368. ISSN: 02643952. URL: http://www.jstor.org/stable/91208 (visited on 01/28/2024).

[26] Ronald A Fisher and Winifred A Mackenzie. "Studies in crop variation. II. The manurial response of different potato varieties". In: *The Journal of Agricultural Science* 13.3 (1923), pp. 311–320.

[27] Vincent François-Lavet et al. "An introduction to deep reinforcement learning". In: *Foundations and Trends® in Machine Learning* 11.3-4 (2018), pp. 219–354.

[28] Milton Friedman. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance". In: *Journal of the american statistical association* 32.200 (1937), pp. 675–701.

[29] Harvey Goldstein and Michael JR Healy. "The graphical presentation of a collection of means". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 158.1 (1995), pp. 175–177.

[30] Danijar Hafner et al. "Mastering atari with discrete world models". In: *arXiv preprint arXiv:2010.02193* (2020).

[31] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[32] Peter Henderson et al. "Deep reinforcement learning that matters". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, 2018. ISBN: 978-1-57735-800-8.

[33] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric statistical methods*. John Wiley & Sons, 2013.

[34] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *NSM3 R library documentation*. https://www.rdocumentation.org/packages/NSM3/versions/1.18. [Online; accessed 03-Apr-2024].

[35] IEEE. *About webpage*. https://www.ieee.org/about/index.html. [Online; accessed 24-Mar-2024].

[36] Riashat Islam et al. "Reproducibility of Benchmarked Deep Reinforcement Learning Tasks for Continuous Control". In: *CoRR* abs/1708.04133 (2017). arXiv: 1708.04133. URL: http://arxiv.org/abs/1708.04133.

[37] Degrave J., Felici F., Buchli J., et al. "Magnetic control of tokamak plasmas through deep reinforcement learning". In: *Nature* (2022), 414–419.

[38] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. "Reinforcement learning: A survey". In: *Journal of artificial intelligence research* 4 (1996), pp. 237–285.

[39] Lukasz Kaiser et al. "Model-based reinforcement learning for atari". In: *arXiv preprint arXiv:1903.00374* (2019).

[40] Ilya Kostrikov, Denis Yarats, and Rob Fergus. "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels". In: *arXiv preprint arXiv:2004.13649* (2020).

[41] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. "Curl: Contrastive unsupervised representations for reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5639–5650.

[42] Sergey Levine et al. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems". In: *arXiv preprint arXiv:2005.01643* (2020).

[43] Nicolai A Lynnerup et al. "A survey on reproducibility by evaluating deep reinforcement learning algorithms on real-world robots". In: *Conference on Robot Learning*. PMLR. 2020, pp. 466–489.

[44] Marlos C Machado et al. "Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents". In: *Journal of Artificial Intelligence Research* 61 (2018), pp. 523–562.

[45] Gregory A Mack and John H Skillings. "A Friedman-type rank test for main effects in a two-factor ANOVA". In: *Journal of the American Statistical Association* 75.372 (1980), pp. 947–951.

[46] Donata Marasini, Piero Quatto, and Enrico Ripamonti. "Inferential confidence intervals for fuzzy analysis of teaching satisfaction". In: *Quality & Quantity* 51 (2017), pp. 1513–1529.

[47] Alberto Martín-Martín et al. "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations". In: *Scientometrics* 126.1 (2021), pp. 871–906.

[48] Encyclopedia of Mathematics. *Student test*. http://encyclopediaofmath.org/index.php?title=Student_test&oldid=54963. [Online; accessed 19-Apr-2024].

[49] Volodymyr Mnih et al. "Asynchronous Methods for Deep Reinforcement Learning". In: *CoRR* abs/1602.01783 (2016). arXiv: 1602.01783. URL: http://arxiv.org/abs/1602.01783.

[50] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *nature* 518.7540 (2015), pp. 529–533.

[51] Peter Bjorn Nemenyi. *Distribution-free multiple comparisons*. Princeton University, 1963.

[52] Steven L Salzberg. "On comparing classifiers: Pitfalls to avoid and a recommended approach". In: *Data mining and knowledge discovery* 1 (1997), pp. 317–328.

[53] Tom Schaul et al. "Prioritized experience replay". In: *arXiv preprint arXiv:1511.05952* (2015).

[54] John Schulman et al. "Proximal policy optimization algorithms". In: *arXiv preprint arXiv:1707.06347* (2017).

[55] Max Schwarzer et al. "Data-efficient reinforcement learning with self-predictive representations". In: *arXiv preprint arXiv:2007.05929* (2020).

[56] Andrew W Senior et al. "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577.7792 (2020), pp. 706–710.

[57] Antonis Sidiropoulos, Dimitrios Katsaros, and Yannis Manolopoulos. "Generalized Hirsch h-index for disclosing latent facts in citation networks". In: *Scientometrics* 72.2 (2007), pp. 253–280.

[58] David Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144.

[59] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587 (2016), pp. 484–489.

[60] John H Skillings and Gregory A Mack. "On the use of a Friedman-type statistic in balanced and unbalanced block designs". In: *Technometrics* 23.2 (1981), pp. 171–177.

[61] Student. "The probable error of a mean". In: *Biometrika* 6.1 (1908), pp. 1–25.

[62] Charles Sutton and Linan Gong. "Popularity of arXiv.org within Computer Science". In: *CoRR* abs/1710.05225 (2017). arXiv: 1710.05225. URL: http://arxiv.org/abs/1710.05225.

[63] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[64] Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Vol. 4. Jan. 2010. DOI: 10.2200/S00268ED1V01Y201005AIM009.

[65] Gerald Tesauro. "TD-Gammon, a self-teaching backgammon program, achieves master-level play". In: *Neural computation* 6.2 (1994), pp. 215–219.

[66] David Tranfield, David Denyer, and Palminder Smart. "Towards a methodology for developing evidence-informed management knowledge by means of systematic review". In: *British journal of management* 14.3 (2003), pp. 207–222.

[67] Warren W Tryon. "Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests." In: *Psychological methods* 6.4 (2001), p. 371.

[68] Warren W Tryon and Charles Lewis. "An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor." In: *Psychological methods* 13.3 (2008), p. 272.

[69] Hado P Van Hasselt, Matteo Hessel, and John Aslanides. "When to use parametric models in reinforcement learning?" In: *Advances in Neural Information Processing Systems* 32 (2019).

[70] Oriol Vinyals et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* 575.7782 (2019), pp. 350–354.

[71] Oriol Vinyals et al. "Starcraft ii: A new challenge for reinforcement learning". In: *arXiv preprint arXiv:1708.04782* (2017).

[72] Martijn Visser, Nees Jan Van Eck, and Ludo Waltman. "Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic". In: *Quantitative science studies* 2.1 (2021), pp. 20–41.

[73] Christopher Watkins. "Learning from Delayed Rewards". PhD thesis. King's College, Oxford, 1989.

[74] Paul J Werbos. "Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research". In: *IEEE Transactions on Systems, Man, and Cybernetics* 17.1 (1987), pp. 7–20.

[75] Karen K Yuen. "The two-sample trimmed t for unequal population variances". In: *Biometrika* 61.1 (1974), pp. 165–170.

[76] Juan Zhang et al. "Comparing keywords plus of WOS and author keywords: A case study of patient adherence research". In: *Journal of the Association for Information Science and Technology* 67.4 (2016), pp. 967–972.